Computational and Experimental Analyses of Promoter Architecture in Yeasts

by

Derek Yung-Ho Chiang

B.S. (University of North Carolina, Chapel Hill) 2000

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Molecular and Cell Biology

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Michael B. Eisen, Chair
Professor Steven E. Brenner
Professor Caroline M. Kane
Professor Jeremy W. Thorner
Professor Mark van der Laan

Spring 2005

The dissertation of Derek Yung-Ho Chiang is approved:

Chair Date

Date

Date

Date

Date

University of California, Berkeley

Spring 2005

Computational and Experimental Analyses of Promoter Architecture in Yeasts

by Derek Yung-Ho Chiang

Abstract

Computational and Experimental Analyses of Promoter Architecture in Yeasts

by

Derek Yung-Ho Chiang

Doctor of Philosophy in Molecular and Cell Biology

University of California, Berkeley

Professor Michael B. Eisen, Chair

Changes in gene expression represent a dynamic response of cells to environmental cues. A fundamental challenge is to understand how regulatory information that specifies gene expression changes is encoded in genome sequences. The multifactorial regulation of eukaryotic transcription initiation is influenced by promoter architecture, which governs the assembly of multiprotein regulatory complexes that contribute to synergistic gene activation. The motivating thrust of this work is to distill key sequence features of promoter architecture and to understand the mechanisms by which these features regulate transcription initiation in yeast.

I describe two computational approaches to associate short DNA sequences with gene expression changes in yeast. Genome-mean expression profiles indicated the regulatory potential of individual sequences by averaging out the effects of multifactorial regulation. In addition, I integrated comparative sequence data into the analysis of gene expression data, based on the expectation that promoter architecture has been phylogenetically conserved. I predicted interactions between pairs of transcription factors using a series of statistical tests to identify pairs of DNA hexamers that were jointly conserved and closely spaced.

To transform these computational observations into mechanistic insights, I developed a synthetic promoter assay to investigate how reporter gene transcription was affected by varying the spacing and sequence between transcription factor binding sites. I applied this assay to characterize promoter architecture constraints on the collaborative recruitment of the coactivator Met4p by the transcription factors Cbf1p and Met31/32p in response to methionine starvation. I found that the order of binding sites was crucial, and that distance constraints on coactivator recruitment were more flexible than those for cooperatively binding transcription factors. Intriguingly, I discovered that certain sequence contexts between the binding sites abolished gene activation.

In conclusion, the incorporation of positional information for multiple transcription factor binding sites vastly improves the accuracy of regulatory predictions. The requirements of promoter architecture may vary, depending on the particular mechanism of transcription factor interactions. In general, close spacing between transcription factor binding sites appears to be necessary, but not sufficient, for multifactorial regulation. Further studies on the key determinants of sequence context would aid the synthetic design of regulatory sequences.

To my loving parents, Thomas and Cecilia

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I thank Mike Eisen for his mentorship and for providing an inspiring environment in which to pursue my research interests. I have greatly appreciated his zeal for discovery, elucidation of key questions, and support of conference presentations. I also offer special thanks for his Oakland Athletics baseball tickets.

My thesis committee has also been extremely helpful in contributing their wide-ranging expertise to my interdisciplinary project. Steven Brenner advised me on the primacy of using biological questions to motivate computational studies (rather than vice versa) since before I arrived at Berkeley. Caroline Kane encouraged me through qualifying exams and helped me transition from the keyboard to the bench. Jeremy Thorner shared practical suggestions from his inexhaustible knowledge of yeast genetics and molecular biology. Mark van der Laan helped to refine my statistical analyses while eagerly learning biology.

My closest collaborator and fellow Canuck, Alan Moses, has infused many enthusiastic insights into our research. I have thoroughly enjoyed our discussions on a wide range of topics, none of which lies outside the scope of his knowledge or interest.

I am grateful to many people who have generously furthered my research training. Several postdoctoral fellows have taught me molecular cloning, yeast genetics and molecular biology, including Audrey Gasch, Justin Fay, David Nix, Paul Spellman, Angela DePace and Needhi Bhalla. Brandon Davies, Adam Martin and Jonathan Wong

provided yeast strains and reagents, as well as several troubleshooting tips. Jasper Rine and Mike Tyers, at the University of Toronto, graciously allowed me to conduct work in their laboratories. Paul Jorgensen and Traci Lee shared reagents and advice that enabled my co-immunoprecipitation experiments. Emily Hare and Brant Peterson have been fun bench companions. I have also benefited from collaborations with several statisticians, including Peter Bickel, Terry Speed, Sunduz Keles, Katerina Kechris, Eric van Zwet, Pat Flaherty and Roded Sharan.

I have been wonderfully supported by the fellowship of my friends: Mike Jaasma, Troy van Voorhis, Christopher and Alia Pan, Jon and Sharon Touryan, Neil and Christina Shenvi, Keith Casner, Justin Hastings, Eric Friedman, Jennifer Lin and Zoe Lui. I am also indebted to the kind concern of Carrie Bare, Barry Wong, Samantha Zylstra, Julien Smith and Mark Labberton. Finally, I thank the graduate student Veritas Christian Fellowship and First Presbyterian Church of Berkeley for providing a spiritual home during the past five years.

CHAPTER 1

INTRODUCTION

PREFACE: DECIPHERING THE EUKARYOTIC *CIS*-REGULATORY CODE

Diversity is a hallmark of life. Each species features distinct morphological and physiological traits that are inherited via genomic DNA. Studies of mutants that lacked a particular trait demonstrated that certain genes were necessary for that trait. Since the set of genes contained in a genome represents the functional repertoire for that organism, genome sequencing projects were initiated to glean insights about how genome content and architecture contribute to the unique biology of different organisms. Remarkably, genome size and the number of genes encoded in an organism's genome do not correlate with phenotypic complexity (Britten and Davidson, 1969). Nonetheless, many genes found in divergent organisms nonetheless share a high degree of sequence similarity. Furthermore, such homologous genes are involved in the development of widely divergent tissues in different organisms (reviewed in Duffy and Perrimon, 1996). These observations suggest that an important contribution to phenotypic diversity may be the precise spatial and temporal regulation of the expression of genes in response to specific environmental conditions. Therefore, a fundamental problem is to understand how this regulatory information is encoded within genome sequences. Our ultimate goal would be to predict the precise conditions under which any given gene is expressed. Not only would these predictions aid the functional annotation of uncharacterized genes, but comparisons between the regulation of homologous genes in different organisms could also suggest how regulatory divergence underlies phenotypic diversity.

The coordinated regulation of multiple genes provides a key towards decoding *cis*- regulatory information. Also called gene batteries, these gene groups often share related functions that confer a particular cellular capability (Britten and Davidson, 1969).

The simultaneous expression of multiple genes is enforced by common *cis*-regulatory sequences that are shared among member genes in a certain gene group. In addition, eukaryotic gene regulation is often multifactorial. Since the activity of many transcription factors often responds to certain environmental conditions, unique combinations of *cis*-regulatory sequences can make a group of genes responsive to a precise set of environmental conditions. The prevalence of multifactorial regulation also enables the integration of multiple signaling pathways and transcription factors upstream of transcription initiation. As a result, complicated patterns of gene expression can arise from the superimposed regulation of multiple signaling pathways.

My dissertation sought to understand how mechanisms of signal integration are governed by promoter architecture, which refers to distance constraints and sequence context among multiple transcription factor binding sites in yeast intergenic regions. In this introductory chapter, I will review previous biological and computational studies on the encoding of *cis*-regulatory information in genome sequences. Signal integration is accomplished, in part, by the assembly of transcription factors into multiprotein regulatory complexes that are necessary for transcription initiation. After briefly describing these complexes, I will discuss how the organization of regulatory sequences can impose geometric constraints on multiprotein complex assembly. I will then present experimental evidence that different mechanisms of transcription factor interactions may impose different constraints on regulatory sequence organization. Next, I will consider how phylogenetic conservation provides a filter for discerning functional features of regulatory sequences. Finally, I will review how computational models of transcription factor binding specificities and regulatory sequence interactions have been used to classify promoters that generate particular patterns of gene expression.

SECTION 1. Regulation of Transcription Initiation in Eukaryotes

Transcription factors transduce cellular signals to alter gene expression

Most transcription factors are sequence-specific DNA-binding proteins with distinct protein domains that can be mixed and matched (Keegan *et al.*, 1986). The DNA-binding domain of a protein makes specific contacts with DNA within a binding site sequence that is typically 5 to 12 nucleotides long. DNA-binding domains belong to a limited set of protein families that recognize groups of similar sequences (Luscombe *et al.*, 2000). Most transcription factors have an activation domain or a repression domain that is dispensable for DNA binding, but necessary to effect gene expression changes (reviewed by Ptashne, 1988). Short peptides within activation domains are sometimes sufficient to recapitulate its function (Ma and Ptashne, 1987a). These peptides tend to have atypical amino acid composition, with stretches of acidic-rich, glutamine-rich or hydrophobic-rich residues. Notably, activation domains from one transcription factor also function in different species, which suggests a conserved mechanism of activation. Indeed, activation domains help nucleate the recruitment of conserved multiprotein regulatory complexes, which modulate the rate of transcription initiation.

Several mechanisms regulate the activity of transcription factors in response to specific environmental conditions. Activation domains can be masked by an inhibitory domain or by interactions with negative regulators. For instance, the Gal4 activation domain is inhibited by Gal80 binding in low galactose conditions (Ma and Ptashne, 1987b). Conformational change of the Hsf1 transcription factor renders it active at higher temperatures (Chen and Parker, 2002). Ligands can also induce conformational changes in several transcription factors, including Put3, Ppr1 and Leu3 (Sze *et al.*, 1992; Flynn and Reece, 1999; Sellick and Reece, 2003). Export to the cytoplasm can prevent transcription factors from binding DNA. For example, phosphorylation of Pho4 mediates its export into the cytoplasm by the exportin, Msn5 (Kaffman *et al.*, 1998).

Enzymology of yeast transcription initiation

Transcription initiation begins after the assembly of RNA polymerase II and many general transcription factors at the transcription start site. General transcription factors in yeast include TFIID, which contains the TATA-box binding protein (TBP or Spt15 in *S. cerevisiae*), TFIIB, TFIIE, TFIIF, and TFIIH (reviewed by Hampsey, 1998, Kuras et al., 2000)). Each of these factors is composed of multiple subunits. Different yeast promoters have different requirements for general transcription factors *in vivo* (Li et al., 2000). The recruitment of TFIID (and thus TBP) is the rate-limiting step of transcription initiation (Ptashne and Gann, 1997; Kuras and Struhl, 1999; Li et al., 1999). TBP recruitment is necessary for transcriptional activation, since mutations were found that block activation by GAL4-VP16, but did not affect basal transcription (Kim et al., 1994). Artificial tethering of TBP to a promoter region via a fusion with the DNA-binding domain of LexA is also sufficient to activate transcription *in vivo* (Chatterjee and Struhl, 1995; Klages and Strubin, 1995).

In eukaryotes, transcription initiation is constitutively repressed by nucleosomes, which consist of approximately 146 bp of DNA wrapped around an octameric complex of the histones H3, H4, H2A and H2B (reviewed by Struhl, 1999). Nucleosome occlusion of binding sites for TBP and transcription factors can be overcome by several mechanisms. First, the N-terminal tails of core histones are accessible to enzymes that regulate covalent modifications of lysine residues (Luger *et al.*, 1997). These modifications serve two purposes. By neutralizing the positive charge on lysine residues, these modifications weaken electrostatic contacts between histones and DNA, thus enabling TBP and other proteins higher accessibility to their binding sites (Anderson *et al.*, 2001). In addition, these modifications serve as novel epitopes for the binding of other regulatory proteins (Stahl and Allis, 2000). Another mechanism for relieving histone-imposed transcriptional repression comes from sliding or displacing nucleosomes in promoter regions, thus increasing the accessibility of transcription factors to DNA (Becker, 2002). Both of these mechanisms combine to clear nucleosomes from the promoter regions of transcribed genes (Boeger *et al.*, 2003, Reinke and Horz, 2003). Indeed, genome-wide studies have reported a strong association between core histone-depleted regions and transcription rates (Bernstein *et al.*, 2004).

A single transcription factor (or its activation domain) can recruit other regulatory factors, including histone acetyltransferases, ATP-dependent chromatin remodeling enzymes, and the mediator complex (Swanson *et al.*, 2003). Mutations in components of these complexes reduce the *in vivo* recruitment of TBP or RNA polymerase at several promoters (Qiu *et al.*, 2004). Chromatin immunoprecipitation studies have also surveyed the order in which regulatory complexes are associated with actively transcribed genes (Gregory *et al.*, 1999; Bhoite *et al.*, 2001; Cosma *et al.*, 2001; Larschan and Winston, 2001; Bryant and Ptashne, 2003). Whereas the order of complex recruitment differs at various promoters, the collaboration of multiple complexes in a multi-step process of gene activation emerges as a common theme. I will now briefly outline some key roles of these complexes.

Histone acetyltransferases

The Gcn5 histone acetyltransferase is the catalytic component of the yeast SAGA complex, which is recruited to the promoters of many genes *in vivo* (reviewed by Carrozza et al., 2003). Numerous studies have proposed a correlation between histone acetylation and transcriptionally active genes (reviewed by Struhl, 1998). However, chromatin immunoprecipitation surveys of histone H3 and histone H4 acetylation in yeast demonstrated that gene activation by various transcription factors is sometimes associated with deacetylation (Deckert and Struhl, 2001). Other yeast histone acetyltransferases include Esa1 – which is recruited to the promoters of ribosomal proteins and heat shock genes (Reid et al., 2000) – as well as Sas2 and Sas3, which oppose silencing at the mating type locus (Kimura et al., 2002; Suka et al., 2002).

ATP-dependent chromatin remodeling complexes

Chromatin remodeling complexes alter the contacts between histones and DNA, leading to nucleosome displacement or exchange (reviewed by Martens and Winston, 2003; Lusser and Kadonaga, 2003). There are several groups of chromatin remodeling enzymes in yeast, including SWI/SNF, RSC, ISWI, and INO80. The Swi2/Snf2 ATPase is the catalytic subunit of the yeast SWI/SNF chromatin remodeling complex. This subunit can be recruited by activation domains of transcription factors, as well as bromodomain-mediated interactions with acetylated lysine residues in histone N-terminal extensions. Microarray studies revealed that the SWI/SNF complex is involved in the regulation of approximately 5% of yeast genes (Sudarsanam *et al.*, 2000). By contrast, the Isw1-containing ISWI complex has been shown to repress transcription of the *PHO8* promoter by displacing TBP (Moreau *et al.*, 2003). These studies suggest that different chromatin remodeling complexes may facilitate either gene activation or repression.

Mediator

The yeast SRB mediator complex bridges interactions between transcription factors and general transcription factors (reviewed by Hampsey and Reinberg, 1999; Myers and Kornberg, 2000). The mediator complex is required for RNA polymerase and general transcription factors to respond to activators in *in vitro* transcription assays. Its multiple roles include TBP recruitment via Gal11 and promoting phosphorylation of RNA polymerase II C-terminal domain by Srb10/Srb11. Chromatin immunoprecipitation studies found that mediator subunits were localized to upstream activating sequences before the recruitment of TBP and general transcription factors (Bhoite *et al.*, 2001; Bryant and Ptashne, 2003; Kuras *et al.*, 2003). These studies further support the model that the mediator complex is recruited before and separately from RNA polymerase II.

SECTION 2. DESIGN PRINCIPLES OF TRANSCRIPTIONAL CONTROL REGIONS

The effects of regulatory sequence organization on gene expression have been long appreciated. Positional organization of transcription factor binding sites can govern protein-protein interactions and thus affect the efficiency of multiprotein complex assembly. In this section, I will review evidence for the enhanceosome model, which proposes that multiprotein complex recruitment requires the stereospecific alignment of transcription factors. Nevertheless, it has become apparent that distance constraints are far from universal. Rather, different types of distance constraints may be enforced, depending on the mechanism of transcription factor interactions. I will discuss some experimental characterizations of helical phasing, precise spacing and short-range distance constraints. However, most of these experiments only tested a handful of different distances with fixed sequences.

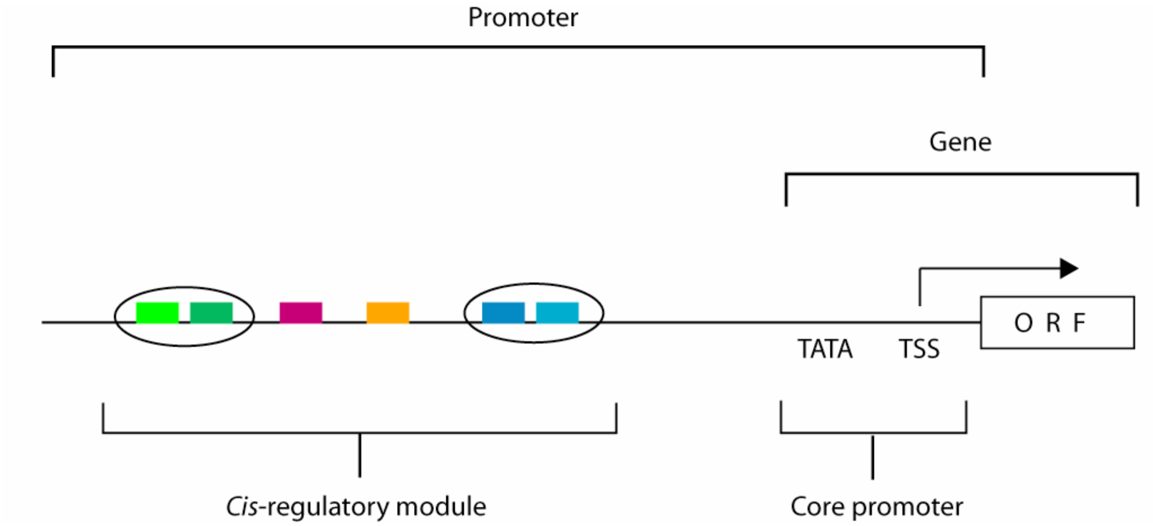
An unresolved question is whether a prototype for distance constraint can be generalized within a family of related transcription factors. Conversely, if each set of interacting transcription factors were to have idiosyncratic positional requirements, it would be much more difficult to derive general principles of promoter architecture. Further structural studies on the interaction surfaces of activation domains with multiprotein complex subunits could enable computational docking predictions on the flexibility of regulatory sequence organization. Despite the wealth of biochemical data on transcription factor interactions, our ability to predict transcription factor interactions from DNA sequences remains woefully incomplete.

Basic anatomy and location of *cis*-regulatory information

Cis-regulatory information can be divided into two main components (Figure 1.1). The core promoter contains certain sequences that position the assembly of the RNA polymerase preinitiation complex. The rate-limiting component for assembly of this complex is the TATA-box binding protein (TBP), which binds approximately 40 to 120 bp upstream of the transcription start site in yeast (Hampsey, 1998). TBP can bind directly to promoters with the consensus sequence TATA(A/T)A(A/T)(A/G); about 20% of yeast promoters contain a match to this consensus that is conserved among alignments of four closely-related yeast species (Basehoar *et al.*, 2004). At other promoters lacking this consensus sequence, TBP can still bind to the core promoter as part of the multiprotein complex TFIID.

Throughout this work, I define a **transcriptional control region** as a DNA sequence that is sufficient to recapitulate a portion of the gene expression pattern of a wild-type gene. These regions often contain several binding sites for multiple transcription factors, which modulate TBP recruitment and thus levels of gene expression (Ptashne and Gann, 1997; Kuras and Struhl, 1999). In yeast, transcriptional control regions are found within several hundred base pairs upstream of the transcription start site. These regulatory regions have also been named upstream activating sequences or upstream repressive sequences (reviewed by Struhl, 1995). Since many yeast upstream activating or repressive sequences are orientation- and distance-independent, they are considered to be functionally analogous to metazoan enhancers (Hampsey, 1998).

Figure 1.1) Hierarchical organization of *cis*-regulatory information



Usually located within several kilobases of a gene, a transcriptional control region comprises a core promoter and one or more *cis*-regulatory modules. The core promoter contains a transcription start site (TSS) and a TATA box. The fundamental unit of regulatory information is a transcription factor binding site, which is indicated by a colored rectangle. Transcription factors may bind cooperatively or antagonistically at composite elements, which are depicted by ovals. *Cis*-regulatory modules comprise multiple binding sites and composite elements.

Cis-regulatory information is hierarchical and modular

The hierarchical organization of *cis*-regulatory information is an important design principle (Figure 1.1). Transcription factor binding sites represent the basic element of regulatory information. The next level of regulatory organization involves the relative placement between two binding sites, which governs protein-protein interactions between transcription factors. Closely-spaced transcription factor binding sites can be considered a functional unit, called a composite element (Diamond *et al.*, 1990). These composite elements implement various schemes of regulatory logic, depending on transcription factor occupancy under different conditions. A key determinant for this combinatorial regulation is the spacing between the binding sites for the individual transcription factors (Pearce *et al.*, 1998). Finally, multiple composite elements can be contained within a *cis*-regulatory module up to several hundred base pairs in length. These modules are functionally defined as the minimal sequence regions that are sufficient to recapitulate a discrete component of normal spatio-temporal regulation, when placed upstream of a reporter gene.

Detection of the coincident binding of two or more different transcription factors can integrate the activation states of multiple upstream signaling pathways. For instance, the yeast sporulation gene, *IME1*, is regulated in response to glucose, acetate, nitrogen and cell type signaling pathways (Vershon and Pierce, 2000). Secondly, the presence of multiple activation domains tethered to a regulatory sequence may achieve synergistic recruitment of downstream enzymatic complexes. In addition, increased target gene specificity can be achieved with binding sites for multiple transcription factors, which contain more information than a single binding site alone. The combinatorial nature of gene regulation helps ensure that inadvertent gene expression does not happen simply due random occurrences of transcription factor binding sites. Finally, various combinations of transcription factor partners provide a modular means of generating regulatory diversity.

Synergistic activation is a consequence of multifactorial transcriptional regulation

Multiple binding sites for the same or different transcription factors can generate a more-than-additive effect on gene activation. For instance, the shared promoter of the divergently transcribed *GAL1* and *GAL10* genes contains four low-affinity binding sites for the transcriptional activator, Gal4 (Giniger & Ptashne, 1988). A single Gal4 binding site only supported 1% of endogenous gene activation in response to galactose. However, two low-affinity binding sites within 45 bp increased reporter gene expression to 20% of the wild-type promoter. This synergistic effect on gene activation has also been observed for multimers of other transcription factors.

Synergistic gene activation could arise from protein-DNA or protein-protein interactions that influence transcription factor assembly on DNA. An increase in the probability of a transcription factor bound to a transcription control region would prolong downstream signals that ultimately recruit RNA polymerase. Protein-protein interactions between the same or different transcription factors can increase the affinity for the ternary complex with DNA, thus increasing occupancy at composite elements (Mueller and Nordheim, 1991). In addition, occupancy of a single transcription factor at a promoter can be increased by multiple copies of its binding site, as demonstrated by methylation protection assays (Giniger and Ptashne, 1988, Tanaka, 1996). This increased occupancy may be caused by an increase in the local concentration of transcription factor or by collaborative removal of nucleosomes from DNA. In the latter case, the binding of a single transcription factor may overcome a large energetic penalty for increasing nucleosome accessibility, making it less energetically costly for subsequent transcription factors to bind (Polach and Widom, 1996; Vashee *et al.*, 1998; Miller and Widom, 2003).

Another model for synergistic gene activation, called the multiple contact model, invokes the recruitment of auxiliary regulators (Herschlag and Johnson, 1993). Multiple bound transcription factors could interact with distinct subunits of various regulatory complexes, such as coactivators, general transcription factors, histone acetyltransferases and nucleosome remodeling enzymes. By stimulating different rate-limiting steps in the progression to transcription initiation, multiple transcription factors can exert a kinetic enhancement of overall transcription.

Promoter architecture governs the assembly of multiprotein regulatory complexes

Polymerase recruitment depends on the geometry and stability of multiprotein complexes

Transcriptional control regions can be considered as DNA scaffolds that bring multiple transcription factors in close proximity, thus enhancing protein-protein interactions that synergistically influence gene expression. Molecular scaffolds implement several design principles that integrate multiple binding events (Dueber *et al.*, 2004). By orienting and increasing the local concentration of the individual transcription factor components, scaffolds facilitate protein-protein interactions between them. These interactions are specified by steric conformations imposed by distance constraints between individual transcription factor binding sites. Scaffolds also detect the coincident binding of multiple proteins. For instance, composite elements enforce the mutual binding of two or three transcription factors to adjacent sites, through increases in binding affinity that involve direct protein-protein interactions between the individual proteins. An alternate mechanism of coincidence detection could be achieved via a protein that interacts with multiple transcription factors. Indeed, separate domains of the coactivator protein CBP/p300 can simultaneously interact with bound transcription factors (Ikeda *et al.*, 2002; Goto *et al.*, 2002).

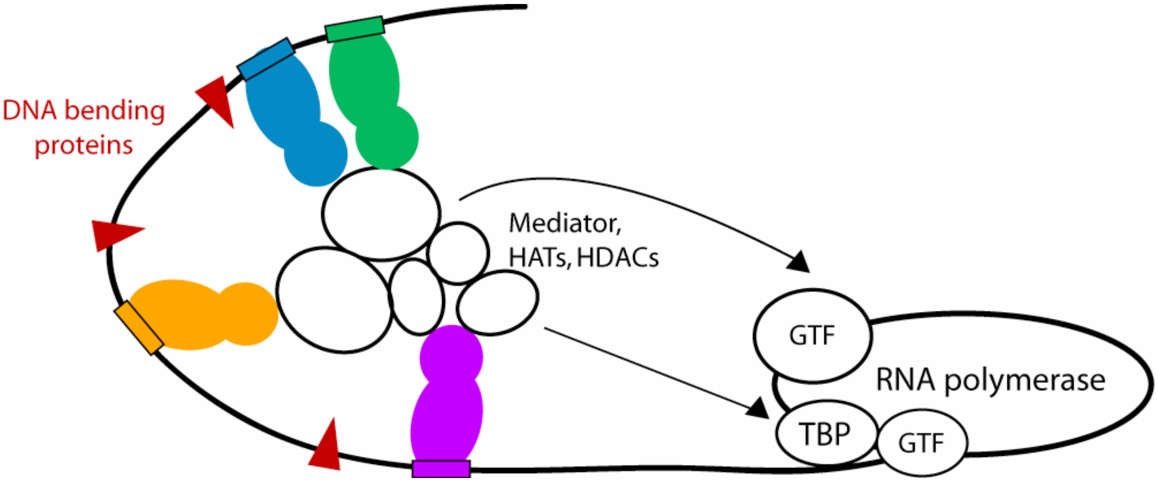
Countless experiments have demonstrated that variations in spacing between transcription factor binding sites can alter the gene expression output specified by a regulatory region. Thus, transcriptional control regions encode both sequence information for a unique combination of transcription factors, as well as positional information between the binding sites. Throughout this work, I will use the term **promoter architecture** to refer to the distance constraints and sequence context among multiple transcription factor binding sites in transcriptional control regions, particularly the intergenic regions of yeast. The arrangement of bound transcription factors often enforces precise geometrical conformations that represent optimal interaction surfaces for downstream coactivators and enzymes. Various combinations of bound activator domains could mix and match to form different interaction surfaces that specifically recruit different coactivators.

<u>Coincident binding of multiple transcription factors can be enforced by enhanceosomes</u>

Enhanceosome assembly provides a paradigm for how specific combinations of appropriately spaced transcription factor binding sites can be detected by the transcriptional machinery of the cell (Figure 1.2). A common regulatory theme emerged from the characterization of several proximal enhancers: a stereospecific configuration of bound transcription factors and their associated activation domains was required for synergistic gene activation (Merika and Thanos, 2001). The relative arrangements of bound transcription factors modified their aggregate protein-protein interaction surface, thus imposing geometric constraints on the nature of the coactivators, modification enzymes, and general transcription factors subsequently recruited.

The enhanceosome model for the integrated effect of multiple transcription factors resulted from studies of the interferon-beta enhancer (reviewed by Grosschedl, 1995; Carey, 1998). This enhancer contains four transcription factor binding sites within 65 bp. *In vitro* reconstitution experiments of enhancer DNA with transcription factors elucidated several major regulatory principles. First, the assembly of transcription factors at the enhancer depended on the orientations of binding sites and the distances between them (Kim and Maniatis, 1997). The native orientation of the ATF-2/Jun site was required for interacting with IRF-1 and recruiting it to the enhancer (Falvo *et al.*, 2000), whereas an insertion of 6 bp between the IRF-1 and NF-kB binding sites reduced gene activation (Kim and Maniatis, 1997). In addition, stereospecificity of protein-protein interactions among activator domains influenced the level of coactivator recruitment. The activation domains of bound transcription factors were not interchangeable, suggesting that a unique surface of activation domains was required for recruitment of the coactivator, CBP/p300 (Merika *et al.*, 1998).

Figure 1.2) Enhanceosomes direct the stereospecific assembly of multiprotein regulatory complexes



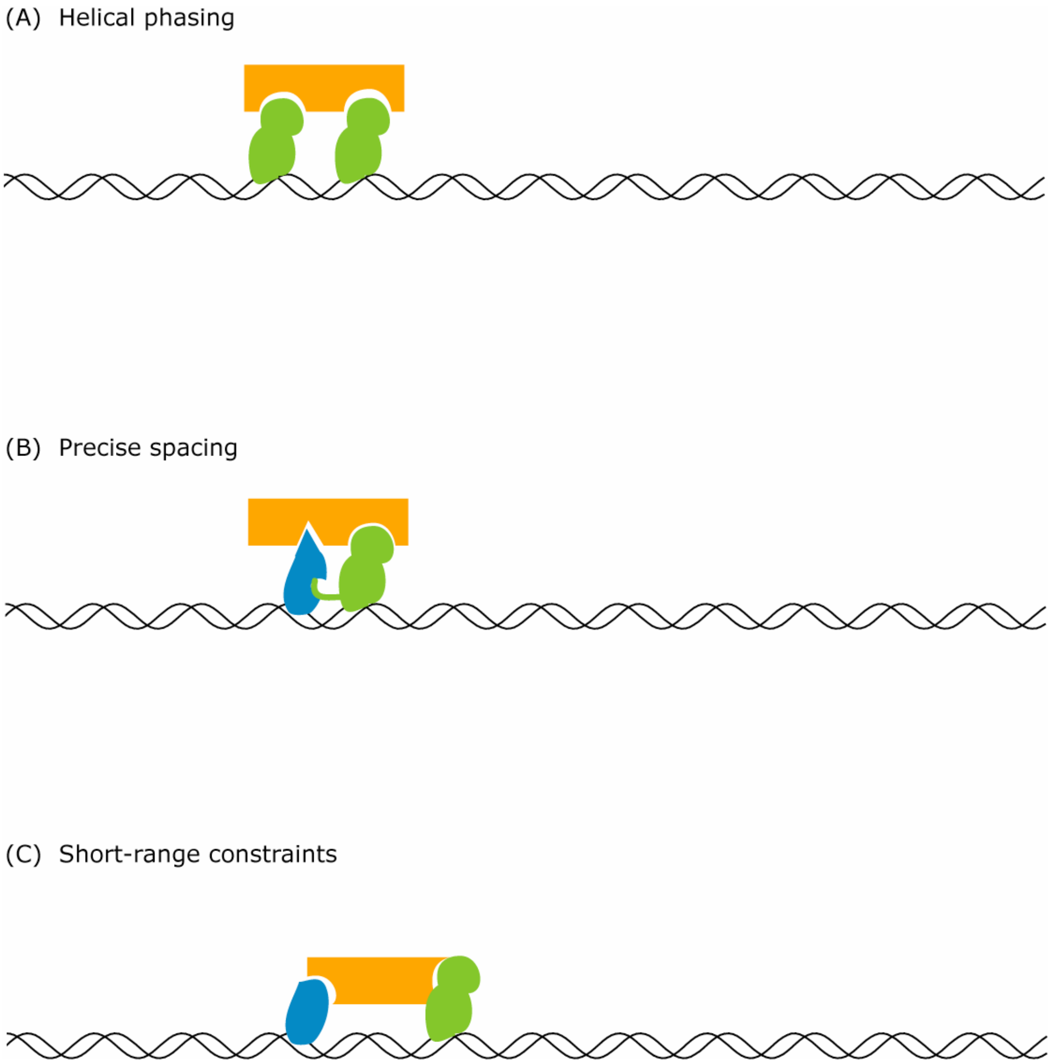
Enhanceosomes represent organized scaffolds for the binding of multiple transcription factors, which are depicted as colored shapes. The precise spatial orientation of transcription factor activation domains directs the synergistic recruitment of multiprotein regulatory complexes, which are indicated by white ovals. These complexes may include histone acetyltransferases, ATP-dependent chromatin remodeling enzymes, and mediator. Stereospecific recruitment imposes constraints on the order, orientation and relative spacing of transcription factor binding sites in enhanceosomes. Architectural proteins, indicated by red triangles, also modify DNA bending.

DNA bendability proved to be another key component of enhanceosome design. Circular permutation assays demonstrated that the high-mobility group transcription factor, HMG I(Y), reverses DNA bending caused by ATF-2 binding, thus favoring interactions between c-Jun and NF-κB (Falvo *et al.*, 1995). *In vitro* reconstitution studies of enhancer transcription demonstrated that HMG I(Y) was a critical component for maximal activation of the interferon-beta enhancer (Kim and Maniatis, 1997). In addition, the HMG I(Y) protein can influence transcription factor assembly directly by interactions with ATF-2 (Kim and Maniatis, 1997). These studies implicate both intrinsic DNA bending and proteins that alter DNA bending as strategies for regulating protein-protein interactions and thus transcriptional activation levels.

Distance constraints may depend on mechanism of transcription factor interactions

The enhanceosome model proposes that the recruitment of multiprotein regulatory and enzymatic complexes requires the precise stereospecific binding of transcription factors in transcriptional control regions. Therefore, the relative placement of transcription factors should affect the extent of multiprotein complex assembly and subsequent gene expression. Several major categories of distance constraints between transcription factor binding sites have been characterized, which correspond to different mechanisms of transcription factor interactions and subsequent recruitment of cofactors (Figure 1.3). In the following section, I will review the major categories of distance constraints.

Figure 1.3) Major categories of distance constraints between transcription factor binding sites



Distance constraints between binding sites often reflect the mechanism by which transcription factors interact. (A) Helical phasing between transcription factors may indicate that activation domains provide a unique recruitment surface on the same face of the DNA helix. (B) Precise spacing is often imposed by direct protein-protein interactions between transcription factors. (C) Short-range distance constraints may reflect an indirect interaction that is bridged by a coactivator.

Helical phasing

The helical phasing between binding sites may be critical for transcription factors that require activation domains on the same face of the DNA helix. In eukaryotic cells, this effect was first characterized in the SV40 early enhancer (Takahashi *et al.*, 1986). Insertions of odd multiples of half a DNA turn (5, 15, 25 bp) between the enhancer and a tandem array of Sp1 binding sites reduced gene activation by up to 90%. In contrast, insertions that preserved helical phasing between these two regions could achieve relatively high levels of gene expression.

At short distances, transcription factors bound to the same face of the DNA helix could present larger interaction surfaces to each other and to other proteins. One consequence of larger surfaces is the facilitation of cooperative binding through favorable energetic interactions. For example, the heat shock transcription factor *HSF* binds cooperatively to a high-affinity site and a low-affinity site in the promoter of the *Drosophila hsp70* gene. Whereas two studies demonstrated that helical phasing between the binding sites modulates reporter gene activation, they conflicted on the maximal spacing between the binding sites (Cohen and Meselson, 1988; Amin *et al.*, 1994). Cohen and Meselson reported that insertions of 30 bp, 70 bp or 300 bp reduced reporter gene expression by only 2-fold, whereas Amin *et al.* observed a 10-fold reduction in reporter gene expression with distances greater than 20 bp. Although the assay sensitivities of Northern blots versus beta-galactosidase staining could account for this difference, it is possible that the nucleotide composition of the spacer sequences also affected cooperative binding. Helical phasing has also been reported between binding sites for two different transcription factors, such as for NF-Y and SRF binding sites in the human beta-actin proximal promoter (Danilition *et al.*, 1991).

Precise spacing

Composite elements demonstrate two distinct types of distance constraints. Some transcription factors only interact under strict conformational requirements. The insertion of a single base pair between binding sites causes a rotational shift of approximately 35°. In some cases, this rotation leads to steric incompatibility and abolishes cooperative binding (Jin *et al.*, 1995; Tan and Richmond, 1998). Since related transcription factors with slightly different interaction geometries can discriminate the distance between binding sites, composite elements with different spacer sizes could impose specificity on transcription factor binding.

Certain transcription factor families exploit this regulatory principle. Within a family, DNA-binding domains recognize similar sequences, yet each transcription factor may have unique protein interaction surfaces that mediate pairing with specific partners. Different members of the nuclear receptor family recognize hexameric half-sites that are found in various orientations and spacings (Remenyi *et al.*, 2004). Interestingly, the retinoid X receptor (RXR) can bind as a heterodimer with four different family members. The spacing between the half-sites (1, 3, 4 or 5 bp) specifies the partner for RXR at a particular binding site. RXR presents different interaction surfaces to its different binding partners at various spacings (Rastinejad *et al.*, 1995). Analogously, homodimers of the zinc binuclear cluster family of transcription factors found in fungi recognize triplet half-sites that are spaced between 0 and 11 bp apart (Akache *et al.*, 2001). Domain swap experiments showed that a 19-residue sequence determines the interacting partner geometry, thus specifying the half-site spacing that is recognized (Reece and Ptashne, 1993). Crystal structures demonstrated that the dimerization domains of various family members interact with distinct geometries, thus providing a mechanism for distinguishing half-site spacings (King *et al.*, 1999).

Short-range distance constraints may govern cooperative binding interactions

Other composite elements can tolerate some flexibility in the distance between their binding sites. Different effects of distance changes on gene expression could be expected depending on the regulatory mechanism. Since protein-protein interactions will experience rotational strain with base pair insertions, increased spacing between binding sites should lower the affinity of transcription factors that bind cooperatively to DNA. Conversely, decreasing the spacing between binding sites may cause steric hindrance, thus occluding one of the transcription factors from binding. Under the mechanism of independent binding, individual transcription factors should have unchanged DNA affinities for binding sites spaced at different distances apart. However, distance changes between binding sites could alter the interaction surface presented by activation domains, thus weakening the recruitment efficiency of the transcription factor pair. In either case, one would expect that the extent of gene activation should decrease as the spacing between binding sites is increased. Surprisingly, this hypothesis has seldom been tested systematically. Most studies examine the effect of fewer than five different distances between transcription factor binding sites on gene activation. With such low sampling, it can be difficult to determine the minimum and maximum distances between transcription factor binding sites for which synergistic activation can occur.

Before my work began, the most detailed characterization of distance constraints between any pair of yeast transcription factors was reported for Rap1 and Gcr1/Gcr2 (Drazinic *et al.*, 1996). Footprinting studies showed that binding sites for the individual transcription factors were 13 bp apart, as measured by the center-to-center distance between the binding sites. Insertions between 5 bp and 30 bp were introduced at 5 bp intervals between the footprinted binding sites in the upstream activating sequence of the *PYK1* gene. Whereas an insertion of 5 bp abolished activation of a *lacZ* reporter gene, an insertion of 10 bp restored activation to half of wild-type levels. Insertions of 15 or more bp showed a monotonic decrease in *lacZ* levels, which were reduced over 8-fold compared to the native spacing. The extent of reporter gene activation correlated with the occupancy of the Gcr1/Gcr2 site, as demonstrated with *in vivo* guanine methylation protection assays. *In vitro* gel shift assays showed that Rap1 recruited Gcr1/Gcr2 to the composite element, and that this recruitment strength decreased monotonically at distances greater than 23 bp. For distances of 23 bp or below, the helical phasing between binding sites governed the extent of synergistic activation.

Promoter architecture: examples from *Saccharomyces cerevisiae*

Common features of promoter architecture impose cell type specificity

The expression of mating-type specific genes is the prototypic example of yeast combinatorial transcriptional regulation (reviewed by Herskowitz, 1989; Johnson, 1995). Yeast exist as one of three cell types: haploid **a**, haploid α, and diploid **a**/α. Cell type is maintained by the coordinated expression of **a**-specific genes, α-specific genes, or haploid-specific genes. Among the cell-type determining genes are the transcription factors a1 (in **a** cells), as well as α1 and α2 (in α cells). The transcription factor Mcm1 is also found in all cell types. Although each of these transcription factors has low sequence specificity, only recognizing 4 unique base pairs, transcription factor pairs are sufficient to regulate the ≤10 genes that are specific to each cell type. Two common features of promoter architecture impose constraints on transcription factor interactions, thus enabling such precise regulation. Composite elements enforce the cooperative binding of transcription factors to adjacent sites, thus facilitating protein-protein interactions. These interactions are also stabilized by DNA bending induced by the transcription factors. Combined genetic, biochemical and structural studies of these transcription factor pairs demonstrate stringent constraints on promoter architectures for different cell types.

Cooperative binding to low affinity sites is a common regulatory strategy at all cell type-specific promoters. For repression in diploids, haploid-specific promoters share a 20 bp composite element that contains binding sites for the repressors, a1 and α2.
At **a**-specific promoters, a high-affinity Mcm1 binding site is flanked on both sides by binding sites for α2. The central Mcm1 binding site imposes orientation and spacing constraints that permit efficient recruitment of α2 (Smith and Johnson, 1992).
At α-specific promoters, the transcription factor Mcm1 binds cooperatively with α1p to a 22 bp composite element, which is necessary and sufficient to activate gene expression (Bender and Sprague, 1987; Inokuchi *et al.*, 1987; Jarvis *et al.*, 1988). Point mutants in Mcm1 that abolish interactions with α1 reduce binding to the composite element *in vitro*

and gene activation *in vivo* (Bruhn and Sprague, 1994). Mcm1 binds cooperatively to α2 via a different interaction surface to repress **a**-specific genes. Thus, the cooperative binding of transcription factor pairs specifies which genes are regulated in response to the transcription factors present in a particular cell type.

The precise spacing between individual binding sites is crucial for cooperative binding at cell type-specific promoters. At haploid-specific promoters, the insertion of a single base pair between the a1 and α2 binding sites abolishes cooperative binding (Jin *et al.*, 1995). At a-specific promoters, insertions of 3 base pairs on either side of the Mcm1 binding site abolishes α2-mediated repression (Smith and Johnson, 1992). Intriguingly, the distance between binding sites varies by a single base pair in **a**-specific promoters (Mead *et al.*, 1996). For these two different spacings, structural studies reveal that a region of α2 that interacts with Mcm1 must adopt different conformations, causing a switch from an alpha helix to a beta-sheet (Tan and Richmond, 1998). Thus, the rotational strain introduced by a single base pair insertion requires a rearrangement of the protein-protein interaction surface to maintain direct contact. The rigidity of spacing constraints selects for a contiguous composite element, thus ensuring that individual binding sites occurring at random locations will not enable cooperative binding.

DNA bending is a common feature of cell type promoter architecture, and may be required for full activation or repression. The binding partners of α2 induce pronounced DNA bends. The crystal structure of the a1-α2 complex bound to DNA revealed an overall bend of 60°, which was absent from the structure of α2 bound alone (Wolberger *et al.*, 1991; Li *et al.*, 1995). Similarly, the crystal structure of Mcm1-α2 bound to DNA showed a bend angle of 72°, which facilitates protein-protein interactions (Tan &

Richmond 1998). Whereas a crystal structure of the DNA-bound Mcm1-α1 complex has not been obtained, circular permutation assays of binding site sequences have shown that this complex of transcription factors bend DNA *in vitro*. Point mutants in Mcm1 with reduced DNA bending fail to form ternary complexes with α1 and DNA *in vitro* and also reduce reporter gene activation over 10-fold *in vivo* (Lim *et al.*, 2003; Carr *et al.*, 2004).

Promoter architecture requirements for co-activator recruitment are poorly understood

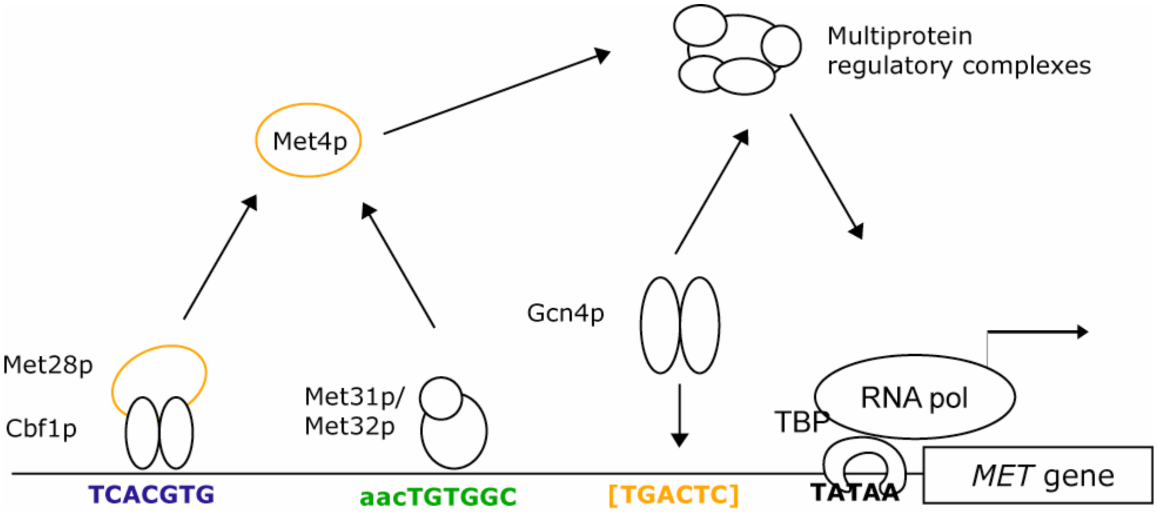
Genetic and biochemical analyses have characterized the transcription factor network that regulates the expression of yeast sulfur utilization genes, which are involved in amino acid metabolism, cell cycle regulation and glutathione-mediated response to oxidative stress (Thomas and Surdin-Kerjan, 1997; Patton *et al.*, 2000; Dormer *et al.*, 2000). The *cbf1*, *met4*, *met28* and *met30* strains, as well as the *met31 met32* double mutant, are methionine auxotrophs (Cai and Davis, 1990; Thomas *et al.*, 1992; Thomas *et al.*, 1995; Blaiseau *et al.*, 1997; Cherest *et al.*, 1997). Deletion analysis of the *MET17* promoter showed that two sequences were necessary for transcriptional activation in response to low intracellular concentrations of *S*-adenosylmethionine (Thomas *et al.*, 1989). These sequences correspond to binding sites for Cbf1 and the paralogs Met31 or Met32, and are also found in the promoters of many methionine biosynthetic enzymes (reviewed by Thomas and Surdin-Kerjan, 1997). However, LexA fusions with Cbf1, Met31 or Met32 could not activate reporter genes placed downstream of LexA binding sites, indicating that these transcription factors lack activation domains (Thomas *et al.*, 1992; Blaiseau *et al.*, 1997). Rather, Met4 lacks a DNA-binding domain, but contains an activation domain, and is required for transcriptional activation downstream of Cbf1 or

Met31 or Met32. Gel shift studies showed that Cbf1-Met28-Met4 could assemble *in vitro* on the upstream activating sequence from the *MET16* promoter (Kuras *et al.*, 1997). In addition, Met31-Met28-Met4 or Met32-Met28-Met4 complexes could assemble on sequences from the *MET3* or *MET28* promoters (Blaiseau and Thomas, 1998). Furthermore, yeast two-hybrid assays with truncation mutants revealed distinct regions of Met4 that mediate interaction with either Cbf1 or Met31 or Met32 (Blaiseau and Thomas, 1998). Taken together, these experiments suggest a model in which the co-activator Met4 is coordinately recruited by the transcription factors Cbf1, Met28, and Met31 or Met32 to the promoters of sulfur utilization genes (Figure 1.4). Nevertheless, the distance constraints between Cbf1 and Met31 or Met32 binding sites have not been delineated.

Several characteristics of sulfur utilization gene regulation render it an ideal model system to study how promoter architecture influences co-activator recruitment. Microarrays have identified a relatively large set of 25 genes that require Met4 for activation under sulfur limitation conditions (Fauchon *et al.*, 2002). Since the binding specificities for the transcription factors Cbf1 and Met31 or Met32 have high information content, high-affinity binding sites can be easily represented by consensus sequences. In addition, transcriptional activation is easy to induce experimentally by withholding methionine from the growth media. Finally, the mechanism for Met4 recruitment has been fairly well characterized experimentally.

A comparison of sulfur and phosphate gene regulation exemplifies how multifactorial regulation can recognize distinct sets of target genes. While both transcription factors are members of the basic helix-loop-helix family of transcription

Figure 1.4) Transcription factor regulatory network that regulates sulfur utilization genes in *Saccharomyces cerevisiae*



Sulfur utilization genes, including methionine biosynthesis enzymes (*MET* genes), are de-repressed in response to low intracellular concentrations of *S*-adenosylmethionine. Two transcription factors collaboratively recruit the coactivator Met4 to these promoters. The basic helix-loop-helix transcription factor, Cbf1, binds as a homodimer to the consensus sequence, TCACGTG. Met28 stabilizes Cbf1 binding to its consensus sequence. In addition, the zinc finger transcription factor, Met31 or Met32, binds to its consensus, AAACTGTGGC, of which the last 6 bp represent the core binding sequence. General control for amino acid starvation by the transcription factor, Gcn4, is superimposed on the regulation of some sulfur utilization genes.

factors, Cbf1 induces a set of 25 genes in response to sulfur limitation, whereas Pho4 induces a set of 18 genes in response to phosphate limitation (Ogawa *et al.*, 2000; Fauchon *et al.*, 2002). The Cbf1 (TCACGTG) and Pho4 ([G/C]CACGTG) recognize different nucleotides immediately 5’ to the core CACGTG recognition sequence (Robinson and Lopes, 2000). However, protein-protein interactions with different partners provide additional sequence information that discriminates the two sets of target genes. As discussed above, Cbf1 interacts with Met31 or Met32 at the promoters of sulfur utilization genes, whereas Pho4 binds cooperatively with Pho2 at the promoters of genes activated by phosphate limitation. Thus, unique binding site combinations and distance constraints between them combine to distinguish sulfur- and phosphate-regulated targets, despite the involvement of transcription factors with overlapping sequence specificities.

SECTION 3. The Value of Phylogenetic Comparisons for Understanding Transcriptional Regulation

Researchers have long exploited sequence conservation to discover functional motifs in both coding and noncoding sequences, and numerous studies have noted that individual transcription factor binding sites are conserved. Phylogenetic comparisons assume that the regulated expression of orthologous genes and the binding specificities of orthologous transcription factors are conserved. Some experimental evidence suggests that these assumptions are valid for the four closely related yeast species used for comparative analyses. A key goal of this work is to investigate whether promoter architecture is also subject to purifying selection. Since interactions between transcription factors are functionally important, I expect that constraints on distances and sequence contexts between their binding sites should also be maintained by purifying selection. In Chapter 3, I formulate explicit statistical tests to discover examples of multifactorial regulation in yeast based on the conservation of promoter architecture.

Phylogenetic footprinting can discover regulatory sequences under purifying selection

Orthologous sequences in different organisms are related by descent from a putative common ancestor, but may have accumulated random mutations from the time of divergence. Mutations within functional regions would be deleterious to the organism and should disappear from the population by purifying selection if they are sufficiently detrimental. An alignment of multiple orthologous sequences should reveal positions with lower mutation rates, thus corresponding to putative functional regions within the sequence (Figure 1.5).

Figure 1.5) Conserved blocks in multiple sequence alignments correspond to putative functional regions

```
Scer    CAGTTGTGGGGCCCGCCCGGCCCAATAG-GTAAAC--T---AAAAT-ACA
Spar    CAGTTGTGGGG-----CCGGCCCAATAAAGTAAAC--T---CAAAT-ACA
Smik    CAGTTGTGGGG----CCCGGCCAAATAAAGTCAAC----ATCGAATAACA
Skud    CAGTTGTGGGG---GCCCGGCCCGATCGAGCAAACAATCCTAAAAACACA
Sbay    CAGTTGTGGGGC-CGTC---------------------------------
        ***********     *

Scer    ATAGAAGGG-GTAC---TGAGTGCACGTGACTTATTTT---TTT-TTTTT
Spar    ATAGAGGGG-GTAC----GAGTGCACGTGACCGCAATT--------TGTT
Smik    ATAGGAGTA-GAA---ACTACTGCACGTGACTCAATTT---CTGGTTTTT
Skud    ATAGGAGCGTGTGCACGCGCGCTCACGTGACTGCAATTGCTCTGG--GGT
Sbay    --------------------GGGCACGTGACCGGGTTTGGTTTGG-----
                               ********     **

Scer    GGTTTTAGGTTTCGCTTTTT-TCA----CCTTTTTCTACTTTCTAACACC
Spar    AGTCTATTTTTTATTTTTTTTCCA----CTTCTCTCTACTTTCTAACACC
Smik    GGC-CTGGGATTCTCTATTTTTCC-CTTCTTCTCTCTGCTTTATAACACC
Skud    GGGGGAGTGTTT-TTTTTCTTTCTTTCTCT-CTCTCTACTTTCCAACACC
Sbay    ---------TTTGGGTTTTT-CC--------------------CGACACC
                  **   * * *  *                      *****

Scer    ACAGTTTTGGGCGGGAAG--CGGAAA-CGCCATAGTT-GTAGGTCACTGG
Spar    ACAGTTTTGGGCGGGAAG--CGGAAAACGCCATAGTT-GTAGGTCACCGG
Smik    ACAGTTTTGGGCGGGAAA--CAAAAACCGCCATAGTT-GAAGGTCACTGG
Skud    ACAGTTTTGCGCCCGAAGACCAAAAAACGCCATAGTT-GAAGGCCGCTGG
Sbay    ACAGTTTTTGGG---------------CGCCATAGTTCGCAAGTCGCAG-
        ********                   ********** * * * * * *

Scer    CG--TGAGTCAAGGCCGGGCAGCCAATGACTAAGAACACGAGGTAACTTG
Spar    CG--TGAGTCAAGGCCGGGCAGCCAATGACTAAGAACGCGAGGTAAATTG
Smik    CGCGTGAGTCAAAGCCGGGCAGCCAATGACTAAGAAAAGGAAGTAAACTG
Skud    CG--TGAGTCAAGGCCGGGCAGCCAATGACTAAGA-CGCGAGCTAAAATG
Sbay    CG--TGAGTCAAGGCTGG-CAGCGAATGACTAAGG-CGCAAG--ACAACG
        **  ******** ** ** ****************     *   *    *

Scer    AATTTAACTATTTATAACCAGTGGTAGTTACGAAGACAAA---TTGTTTT
Spar    GATTTAACTATTTATAATCAGTTATAGTTATGAAAACAAG---CCATTTT
Smik    GATTGAACTATTTATAATCGGTTGCAGTTACAGAGAAAGA---TCCTTTT
Skud    GGTTTGACTATTTATAATCGGCGGTAGTTACGAAGACAAGCGCTTCTGTT
Sbay    GATTCGAGTATTTATAATCGGTGGTGGTTACGGGACGAGG---GCGGTTT
          **  * ********* * *      ****       *         **
```

Figure 1.5 (continued)

A multiple sequence alignment for the intergenic region upstream of the *MET28* gene was generated with T-COFFEE (Notredame *et al.*, 2000). The alignment comprises orthologous sequences from *S. cerevisiae* (Scer), *S. paradoxus* (Spar), *S. mikatae* (Smik), *S. kudriavzevii* (Skud) and *S. bayanus* (Sbay). Due to the optimal evolutionary distance of this alignment, contiguous blocks of conserved sequences largely match consensus sequences for transcription factor binding sites: Met31 or Met32 (green, AA[C|T]TGTGG); Cbf1 (blue, CACGTGA); Gcn4 (orange, TGA[C|G]TCA); Yap1 (red, TGACTAA); and TATA box (black, TATAA).

Comparative analyses of protein sequences often identified conserved domains and residues that were necessary for protein function, such as catalysis, DNA binding, or phosphorylation (Vogel *et al.*, 2004). Conserved regions within noncoding sequences could represent a variety of functional elements, including transcription factor binding sites, noncoding RNAs and matrix attachment regions (reviewed by Duret and Bucher, 1997; Cooper and Sidow, 2003).

Phylogenetic footprinting approaches were first applied to orthologous regulatory regions within mammalian genomes (Tagle *et al.*, 1988). This term invokes the principle of experimental footprinting assays, in which a bound transcription factor protects its recognition site from enzymatic modification. Analogously, the process of purifying selection is thought to prevent deleterious mutations from accumulating within a functional element. The predictive power of phylogenetic footprints in identifying transcription factor binding sites was first tested for the γ-globin gene (Gumucio *et al.*, 1996). Among 13 phylogenetic footprints identified, 12 could be bound by specific transcription factors *in vitro*, whereas only 2 of 9 non-conserved regions showed this property. Similar studies have been conducted using comparisons between human and mouse orthologs (Loots *et al.*, 2000) or even among human, mouse and pufferfish orthologs (Aparicio *et al.*, 1995).

Key assumptions and limitations of phylogenetic footprinting

The simplest implementations of phylogenetic footprinting define functional regions as contiguous blocks of aligned residues that are perfectly identical among all species. Software tools such as VISTA and PipMaker generate visual representations of sequence conservation from a user-specified alignment (Loots *et al.*, 2002; Schwartz *et al.*, 2000). Nevertheless, these simple methods share a fundamental caveat: conserved residues within an alignment are not necessarily under purifying selection. Instead, these conserved residues represent a mixture of residues under selection, as well as residues with shared descent that have had insufficient time for mutations to accumulate.

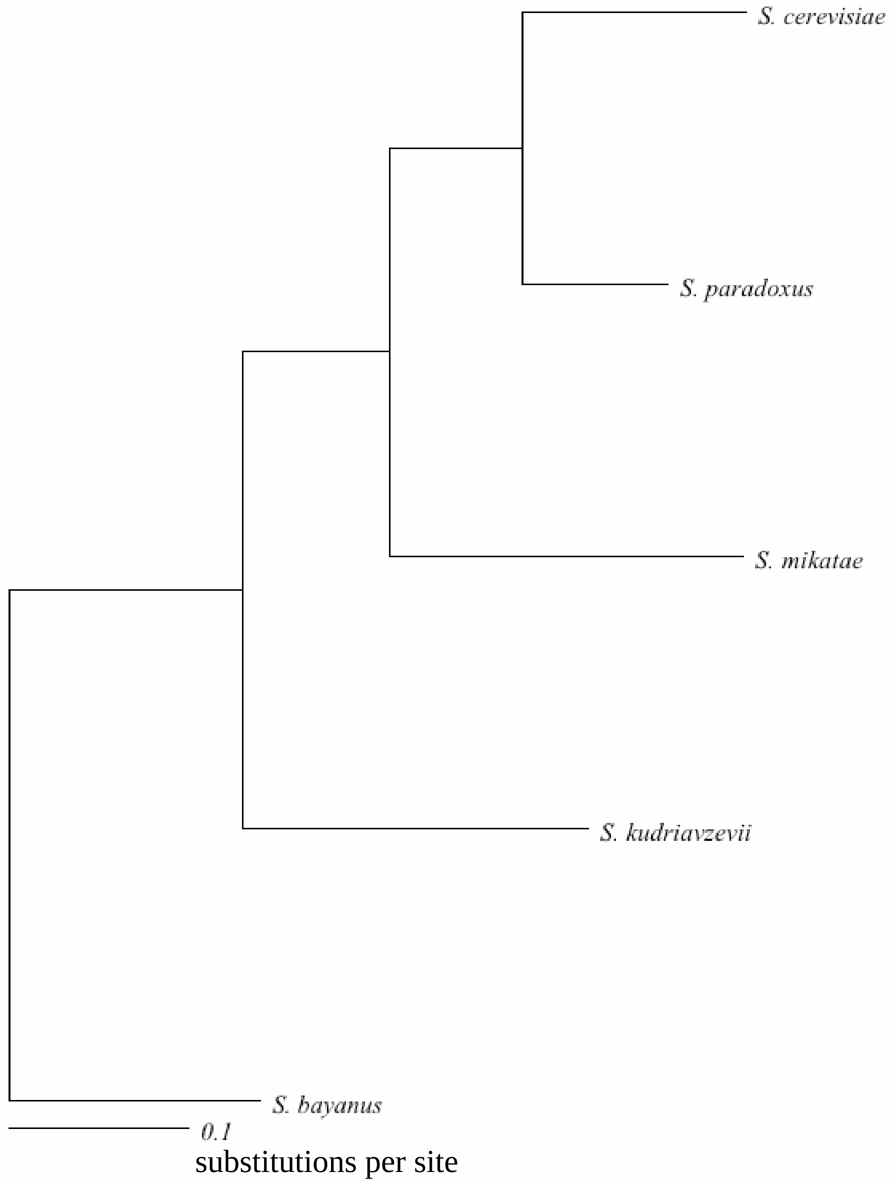
Two key parameters can affect the efficacy of phylogenetic footprinting approaches. First, the sequence alignment must be correct, meaning that residues within each column of the alignment are homologous among the different species. Secondly, the species tree must span sufficient evolutionary distance for random mutations to accumulate within the nonfunctional positions (Cooper and Sidow, 2003). Pairwise alignments of closely related species will not have sufficiently diverged for enough to identify residues under purifying selection. However, sequences that are too divergent are notoriously difficult to align correctly, due to the high frequency of insertions, deletions and inversions within noncoding DNA (Pollard *et al.*, 2004).

Phylogenetic footprinting of transcriptional control regions makes two key assumptions about selection pressures on gene expression. These assumptions are assumed to be true for closely related species, but several examples demonstrate that these assumptions may not be valid at longer evolutionary distances. First, phylogenetic footprinting assumes that orthologous genes retain similar patterns of transcriptional regulation in response to the same environmental conditions. However, microarray studies of mating in *Candida albicans* and cell cycle regulation in *Schizosaccharomyces pombe* have revealed considerable differences in gene expression with orthologs in *S. cerevisiae* (Tsong *et al.*, 2003; Rustici *et al.*, 2004). Second, phylogenetic footprinting assumes that the sequence specificities of transcription factor orthologs are relatively invariant among the species under consideration. Yet transcription factor orthologs can sufficiently diverge in their DNA recognition domains such that their binding specificities are slightly altered. For instance, orthologs of the zinc finger transcription factor, Rpn4, from *Candida albicans* and *S. cerevisiae* showed different binding specificities *in vitro* (Gasch *et al.*, 2004).

The genome sequences of several ascomycete fungi have been recently completed

The availability of complete genome sequences enabled systematic investigations of purifying selection on yeast regulatory sequences. Genome sequences from four closely related *Saccharomyces sensu stricto* were recently completed: *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, and *S. bayanus* (Figure 1.6) (Cliften *et al.*, 2003; Kellis *et al.*, 2003). These four species are similar enough to *S. cerevisiae* that they have a similar karyotype and can form stable diploids (Cliften *et al.*, 2001). Since these species are physiologically similar, it is assumed that the same selection pressures have operated on the regulatory sequences of orthologous promoters. Nevertheless, some differences in nutritional capabilities are observed among these species that could correspond to different selective pressures imposed by their respective ecological niches. Most notably, *S. kudriavzevii* is unable to utilize galactose, since mutations have inactivated all genes of the *GAL* pathway (Hittinger *et al.*, 2004).

Figure 1.6) Phylogenetic tree of *Saccharomyces sensu stricto* species



Phylogenetic comparisons in this work focus on five closely-related *Saccharomyces* species, which can form stable diploids: *S. cerevisiae, S. paradoxus, S. mikatae*, *S. kudriavzevii* and *S. bayanus*. Branch lengths represent the median number of substitutions per site in alignments of intergenic regions, inferred by PAML.

Alignments of the *sensu stricto* yeast species spanned a satisfactory evolutionary distance for phylogenetic footprinting. Pairwise alignments of intergenic regions with *S. cerevisiae* averaged between 60% and 71% identity at the nucleotide level, depending on the species (Leonid Teytelman, UC Berkeley, personal communication). The vast majority of multiple sequence alignments for intergenic regions were high of quality, with contiguous regions of identical columns that could be easily distinguished from mutations within flanking sequences. Due to the ease of identifying conserved sequence blocks, as well as the relative large evolutionary distance spanned by the multiple species tree, I used a simple implementation of phylogenetic footprinting to search for conserved sequence elements. Separate work identified over 80 consensus sequences – some of them corresponding to known transcription factor specificities – by searching for common conserved footprints in related groups of genes (Cliften *et al.*, 2003; Kellis *et al.*, 2003). This success in discovering known consensus sequences strongly implies that some transcription factor binding specificities have been maintained across these yeast species.

High coverage shotgun sequencing has also been conducted on ascomycete fungi as close as *Saccharomyces castellii* and as distant as *Schizosaccharomyces pombe* (see references in Gasch *et al.*, 2004). However, all of these species are too distant to obtain reliable pairwise noncoding DNA alignments with *S. cerevisiae*. Several of these species also exhibited some major morphological and physiological differences, which may correspond with underlying changes in regulatory sequences (Gasch *et al.*, 2004). I thus focused my analyses on the five closely-related *sensu stricto* species.

SECTION 4. COMPUTATIONAL MODELS OF CIS-REGULATORY INFORMATION

A fundamental computational challenge is to predict groups of genes that are expressed under specific environmental conditions. Computational models must incorporate relevant transcription factors, as well as promoter architectures that govern interactions between them. This section will begin with sequence models for transcription factor binding sites, which are the basic unit of regulatory information. Next, I will provide an overview of the formalism of information theory, which quantifies the binding specificity of transcription factors. Finally, I will discuss the strengths and weaknesses of various computational approaches for promoter classification.

Since the vast majority of predicted binding sites are not associated with control of gene expression, additional information must be required to distinguish the physiological targets of a given transcription factor. Some of this information is provided by binding sites for multiple transcription factors. Thus, I propose promoter architecture as a framework for dissecting distance constraints that govern multifactorial control. Whereas previous computational analyses have used heuristic distance cutoffs, few general formalisms have been developed to account for the regulatory logic encoded by multiple transcription factor binding sites (Buchler et al 2003; Istrail & Davidson, 2005). In Chapter 3, I will present a statistical method to evaluate whether conserved binding sites are closer than expected by chance. My experimental characterization of distance constraints between a specific pair of yeast transcription factors will be discussed in Chapter 4. By incorporating organizational principles of regulatory sequences, I aim to glean insights into the underlying biological process of transcription initiation.

Sequence models for transcription factor binding specificities

Transcription factors specifically recognize similar DNA sequences

The DNA-binding domain of a given transcription factor typically recognizes a range of similar sequences. This recognition is usually mediated by hydrogen bonds and van der Waals contacts between individual nucleotides and amino acid side chains of transcription factors. Crystal structures showed that only a subset of nucleotides within a binding site interact specifically with a transcription factor; at other positions, different nucleotides can be tolerated (Luscombe *et al.*, 2000). Strikingly, positions with lower nucleotide degeneracy correlate with increased hydrogen bonding and van der Waals contacts with transcription factor side chains (Mirny and Gelfand, 2002). Binding sites that are recognized by the same transcription factor often show extensive variability, sometimes even within nucleotides that make direct contacts with transcription factor side chains. This variability may alter binding site affinity, thus enabling differential regulation at various promoters. For instance, high-affinity binding sites would have higher levels of transcription factor occupancy than low-affinity binding sites. By varying binding site affinities, different transcription control regions could be activated at different levels of a transcription factor gradient, as with the case for the dorsal-ventral patterning in *Drosophila melanogaster* (Rusch and Levine, 1996).

Experimental methods for compiling transcription factor binding sites or co-regulated genes

Several experimental techniques can be used to compile binding sites for a particular transcription factor. DNase I footprinting exploits the fact that a bound transcription factor protects nucleotides that are buried in the protein-DNA interface from modification by other proteins (Galas and Schmitz, 1978). When incubated with a transcription factor, nucleotides that are resistant to cleavage by partial DNase I treatment thus represent a binding site for that transcription factor. Another method to assay protein-DNA interactions *in vitro* is an electrophoretic mobility shift assay (Garner and Revzin, 1981). Since protein-DNA complexes have a higher molecular weight than free DNA, radiolabeled oligonucleotides containing a binding site will migrate more slowly through a polyacrylamide gel when pre-incubated with a transcription factor. The binding affinities of different sequence variants can be compared by titrating out protein-DNA complex formation with increasing amounts of unlabeled DNA. Finally, binding specificities of transcription factors can be ascertained by *in vitro* selection (Oliphant *et al.*, 1989; Pollock and Treisman, 1990). Pools of random oligonucleotides can be passed over a transcription factor affinity column. Sequences that bind to the column with high affinity can be eluted and recycled through the column. After several rounds of selection, the resulting DNA sequences represent high-affinity binding sites for the transcription factor. Transcription factor binding specificities from these various assays have been compiled in several online databases, such as TRANSFAC (Matys *et al.*, 2003), SCPD (Zhu and Zhang, 1999) and JASPAR (Sandelin *et al.*, 2004).

Global gene expression profiling experiments can identify genes that are likely to be co-regulated, but do not precisely map transcription factor binding sites. For example, high-throughput measurements of gene expression can be obtained by RNA extraction, fluorescent labeling of reverse transcripts and hybridization to an array of oligonucleotide probes that are specific for different genes or genomic regions (Brown and Botstein, 1999). Hierarchical clustering can reveal groups of genes with highly similar expression patterns across multiple treatments (Eisen et al, 1998). Assuming that co-expressed genes are regulated by a common set of transcription factors, binding sites for those proteins are expected to be found in the relevant promoters or enhancers associated with those genes. Another method to detect co-regulated genes is chromatin immunoprecipitation (Taverner *et al.*, 2004). In this method, transcription factors are cross-linked to DNA *in vivo* and isolated. Chromatin is sheared and antibody immunoprecipitation is used to recover DNA fragments bound by the transcription factor. After crosslink reversal and PCR amplification, hybridization of DNA fragments to a microarray spotted with intergenic DNA identifies genomic regions that are bound by a transcription factor *in vivo*. Both of these methods generate a list of putatively regulated sequences that are predicted to contain binding sites for one or more transcription factors. The goal of computational methods is to predict both the sequences and start positions of these binding sites.

Computational models of transcription factor binding specificities

A consensus binding specificity of a transcription factor can be inferred from a multiple sequence alignment of its mapped and putative binding sites. Consensus sequences are the simplest representations of transcription factor sequence preference (Figure 1.7). At each position, the consensus sequence comprises the most frequent nucleotide(s) found in the alignment. The most stringent positions contain one nucleotide, whereas more degenerate positions allow multiple nucleotides. Although consensus sequences are easy to generate, they have limited quantitative value in scoring

Figure 1.7) Representations of transcription factor binding specificity

```
GTGAGTCAC
CTGAGTCAT
ATGAGTCAT
ATGAGTCAC
ATGACTCAT
ATGAGTCAA
ATGACTCAT
GTGAGTCAT
ATGAGTCAT
TTGACTCAT
ATGACTCAT
ATGAGTCAT
ATGAGTCAT
CTGACTCAT
GTGACTCAT
GTGAGTCAT
ATGAGTCAT
GTGACTCAC
ATTAGTCAT
CTGACTCAG
```

(A) Frequency matrix

A	11	0	0	20	0	0	0	20	1
C	3	0	0	0	8	0	20	0	3
G	5	0	19	0	12	0	0	0	1
T	1	20	1	0	0	20	0	0	15

(B) Position weight matrix (log-odds matrix)

A	**0.7**	-2.9	-2.9	**1.5**	-2.9	-2.9	-2.9	**1.5**	-1.9
C	-0.2	-2.2	-2.2	-2.2	**1.0**	-2.2	**2.2**	-2.2	-0.2
G	**0.4**	-2.2	**2.1**	-2.2	**1.5**	-2.2	-2.2	-2.2	-1.2
T	-1.9	**1.5**	-1.9	-2.9	-2.9	**1.5**	-2.9	-2.9	**1.1**
I_{seq}	0.2	1.0	1.4	1.0	1.0	1.0	1.6	1.0	0.5

$$W_{b,i} = \log_2 \frac{f_{b,i}}{p_b} \qquad (1)$$

Consensus sequence:

```
ATGACTCAT
G   G   C
```

$$I_{seq}(i) = \sum_b f_{b,i} \log_2 \frac{f_{b,i}}{p_b} \qquad (2)$$

A multiple sequence alignment of 20 binding sites is displayed at left. The most frequent nucleotides at each position comprise the consensus sequence. (A) Nucleotide counts can also be represented in a frequency matrix. (B) A frequency matrix can be converted to a position weight matrix using equation (1), where $f_{b,i}$ represents the frequency of nucleotide *b* in column *i*, and p_b represents the genome frequency of nucleotide *b*. In the *S. cerevisiae* genome, $p_A = p_T = 0.31$ and $p_C = p_G = 0.19$. To avoid taking logarithms of zero, one pseudocount was added to each cell of the frequency matrix. The information content at each position (I_{seq}) is given by equation (2).

new sequences as putative binding sites. Intuitively, mismatches in conserved positions should be less tolerated than mismatches at degenerate positions. However, consensus sequences provide no formalism for distinguishing the relative effects of mismatches using consensus sequences.

Position weight matrices provide a better representation of binding energies between transcription factors and DNA. Instead of a rigid consensus sequence, each column of a position weight matrix models the conditional probability distribution of finding an adenine, cytosine, guanine or thymine at each position of the binding site alignment (reviewed by Stormo, 2000). These probabilities are converted to log-odds ratios in order to make scores from different positions additive, as well as to normalize against the overall base composition of a given genome (Figure 1.7**)**. Statistical mechanics theory links these log-odds ratios with the contribution of individual nucleotides to binding energies (Berg and von Hippel, 1987). Thus, the affinity of any particular sequence for a transcription factor can be scored as the sum of binding energies at each nucleotide position. Position weight matrices also make several simplifying assumptions that might not always be valid: different positions contribute additively to binding energy; background nucleotide frequencies are independent and identically distributed; all binding sites are equally accessible; and transcription factor binding to DNA is at equilibrium *in vivo*.

Computational discovery of potential transcription factor binding sites

The consensus sequence representation provides a simple way to predict sequences of transcription factor binding sites. The general principle is to enumerate all possible DNA "words" of a fixed length; for instance, there are 4096 (4^6) unique DNA hexamers. For each individual word, its observed frequency in the set of regulated promoters can be compared with its expected frequency, assuming that genome sequences are random. Binomial or hypergeometric tests can evaluate whether a particular word is statistically enriched in the regulated sequences. These enriched words represent putative transcription factor binding sites. Furthermore, highly related words can overlap and thus be compiled into consensus sequences, which often correspond to known transcription factor binding specificities. Several groups have used these approaches to identify binding sites in the promoters of co-regulated yeast genes (van Helden *et al.*, 1998; Brazma *et al.*, 1998; Sinha and Tompa, 2000).

Position weight matrix representations of enriched sequences can also be deduced by computational methods. A position weight matrix of length L consists of $4 \times L$ parameters, *i.e.* the negative log-odds of each nucleotide at each position. These parameters can be optimized using an iterative refinement procedure. Initially, randomly chosen start positions are used to build a position weight matrix. In each subsequent round, these methods calculate the posterior probability that each position in a regulated sequence represents the start of a transcription factor binding site. These probabilities are used to select a new instance of a binding site, either deterministically or stochastically, and the position weight matrix parameters are updated. Over multiple iterations, there is a chance that an actual binding site will be incorporated in the position weight matrix model, thus biasing future search rounds for that transcription factor. These algorithms usually converge on a local minimum that corresponds to an enriched sequence motif. Some popular implementations of this approach include MEME (Bailey and Elkan, 1995), GibbsDNA (Lawrence *et al.*, 1993), AlignACE (Hughes *et al.*, 2000a), and BioProspector (Liu *et al.*, 2001). The first method uses an expectation-maximization approach, whereas the remaining methods are variants of stochastic Gibbs sampling.

Exploiting comparative genomics for motif discovery

Comparative genomics can provide a filter for discovering sequence motifs that are more likely to represent transcription factor binding sites, since evolutionary selection for retention of function should reduce the number of mutations accumulating within sites that are physiologically important *in vivo* (Moses *et al.*, 2003). Several methods use evolutionary models to infer the rate of sequence evolution at all positions in a multiple sequence alignment of orthologous regulatory regions (Boffelli *et al.*, 2003; Wang and Stormo, 2003; Moses *et al.*, 2004; Prakash *et al.*, 2004). These methods search for sequences that show enrichment among a set of input sequences, as well as lower mutation rates than the flanking sequence context within multiple alignments. By explicitly modeling the common ancestry of orthologous sequences, these methods have been shown to perform better than motif discovery algorithms developed for single genomes.

Discriminating target genes regulated by transcription factors

Information content quantifies the contribution of individual positions to binding specificity

The information content at a particular position in a binding site predicts its relative contribution to transcription factor specificity. Information content is inversely related to the statistical entropy at a position (Stormo, 2000):

$$I_{seq}(i) = \sum_{b} f_{b,i} \log_2 \frac{f_{b,i}}{p_b}$$

where $f_{b,i}$ represents the frequency of nucleotide b at position i in the position weight matrix, and p_b represents the frequency of nucleotide b in the whole genome. Positions with zero information content have equal nucleotide probabilities, whereas positions with the highest information content only allow one nucleotide, which indicates specific recognition between that nucleotide and a transcription factor side chain (Mirny and Gelfand, 2002).

An interesting connection between information content and evolutionary constraints based on phylogenetic comparisons was recently demonstrated in yeast (Moses *et al.*, 2003). Experimentally characterized transcription factor binding sites were compiled from the SCPD database (Zhu and Zhang, 1999) and used to construct position weight matrices. These matrices thus reflect the within-genome variability at any position within a binding site. Orthologs to these annotated binding sites were then extracted from multiple sequence alignments constructed from three additional *Saccharomyces sensu stricto* species (Kellis *et al.*, 2003). At each position in a binding site, the minimal number of sequence changes along the species tree was calculated by parsimony (Moses *et al.*, 2003). Strikingly, the rate of substitutions at a particular position was inversely correlated with the information content at that position. In addition, positions with high information content and few evolutionary substitutions were often contacted directly by transcription factor side chains. Assuming that the transcription factor sequence specificities are unchanged over these short evolutionary distances, this analysis implies that similar selection pressures govern the within-genome and between-genome variability of transcription factor binding sites.

Physiological targets of a transcription factor differ from predicted high-affinity binding sites

The total information content of a position weight matrix can predict the frequency of binding site occurrences in the entire genome, under a model that all sequences occur with equal frequencies. Only a small fraction (usually less than 20%) of these predicted binding sites show evidence of transcription factor binding and subsequent changes to gene expression when tested experimentally. Several associations between genomic datasets and predicted binding sites in transcriptional control regions support this observation. First, microarray analyses revealed that the vast majority of computationally predicted binding sites in yeast promoters do not correspond with expected changes in gene expression. For example, the basic helix-loop-helix transcription factor, Pho4, activates gene expression in response to low phosphate conditions. At least one perfect match to the Pho4 consensus sequence, CACGTG, can be found in 376 yeast promoters, defined as the 600 bp upstream of translation start. However, only 12 genes are consistently activated under low phosphate conditions, or about 3% of all genes with at least one predicted Pho4 binding sites (Figure 1.8)

Figure 1.8) Physiologically regulated targets represent a subset of predicted transcription factor binding sites

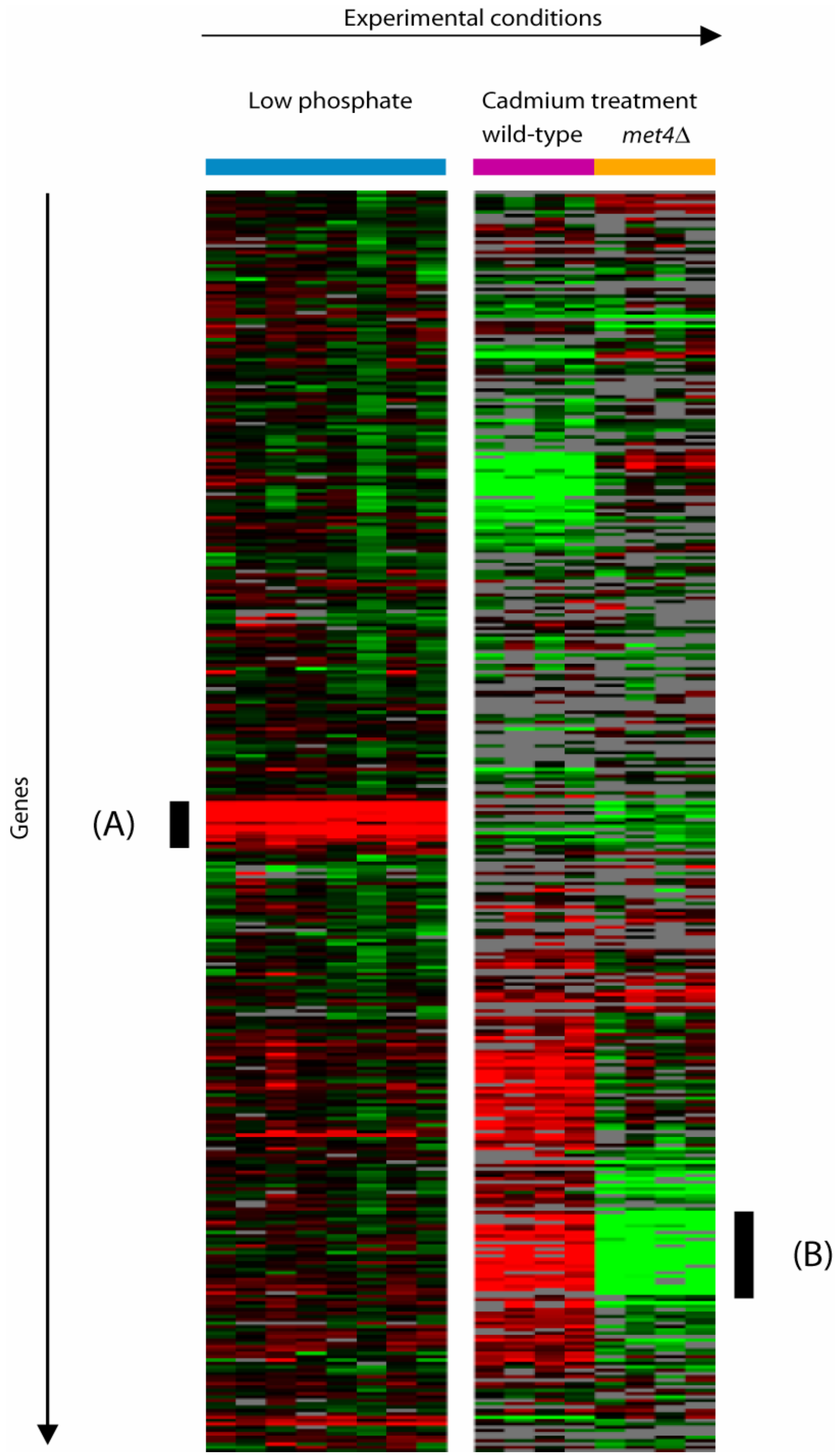


Figure 1.8 (continued)

Gene expression data is displayed for 376 genes that contain an E box motif (CACGTG) within 600 bp upstream of their translation start sites. Each row represents a different gene, whereas each column represents a different microarray that probed low phosphate conditions (Ogawa *et al.*, 2000) or cadmium treatment in wild-type and *met4* deletion strains (Fauchon *et al.*, 2002). Green pixels correspond to transcriptional repression, red pixels correspond to transcriptional induction, and the pixel intensity reflects the magnitude of the change in average gene expression. (A) A subset of 12 genes are induced in response to low phosphate by the transcription factor Pho4. (B) A subset of 26 genes are induced in response to cadmium treatment in a Met4-dependent manner by the transcription factor Cbf1.

(Ogawa *et al.*, 2000). Second, systematic chromatin immunoprecipitation studies have revealed that less than 20% of predicted binding sites are occupied by the corresponding transcription factor *in vivo* (Iyer *et al.*, 2001). Taken together, these data suggest that transcription factors bind a small subset of computationally predicted sites *in vivo*.

Several mechanisms may account for the higher selectivity of *in vivo* targets, compared with predicted *in vitro* binding specificities. Cooperative binding or synergistic activation may require nearby binding sites for two or more transcription factors. Therefore, promoters that contain a predicted binding site for only one transcription factor may not bind that protein or stimulate the recruitment of multiprotein regulatory complexes as efficiently. For instance, activation of the *PHO5* promoter requires adjacent binding sites for the transcription factors, Pho4 and Pho2 (Barbaric *et al.*, 1998). This promoter also exemplifies another mechanism for *in vivo* selectivity: the protection of high-affinity binding sites by nucleosomes (reviewed by Svaren and Horz, 1997, Struhl, 1999). Due to our limited knowledge of nucleosome positioning sequences, it is difficult to computationally predict whether putative transcription factor binding sites are buried or accessible *in vivo*. Another mechanism for *in vivo* selectivity of binding sites may involve local concentration gradients of transcription factors. In this model, several low-affinity binding sites should be found in the vicinity of a high-affinity binding site that is occupied *in vivo*. These low-affinity sites would increase the local concentration of a transcription factor as it scans DNA, thus increasing its probability of retention at that locus.

Promoter architecture can be inferred from an input set of co-expressed genes

Recognizing sequence patterns in the *cis*-regulatory code

A key problem is to decipher how the elements and organization of transcriptional control regions specify changes in gene expression in response to particular environments or cell types. In other words, we want to discover the particular promoter architecture that is sufficient to recapitulate a given pattern of gene expression. Promoter architecture can be inferred by statistical analyses of promoter sequences and co-regulated genes. The input to these algorithms includes a list of co-regulated genes, as well as a list of transcription factor binding specificities. These lists can be specified in advance or predicted by the algorithm. Each algorithm adopts its own definition of promoter architecture, which usually comprises a list of transcription factors, the relative affinities of their binding sites and the distance constraints between binding sites. These parameters are optimized to include all examples from the training set of co-regulated genes. Various algorithms use different criteria to optimize models of promoter architecture. Template methods extrapolate sequence features from a small set of transcriptional control regions with experimentally mapped transcription factor binding sites. Structured motif discovery methods search for sequence pairs and distance constraints that discriminate a user-specified set of co-regulated genes from the rest of the genome. The resulting model of promoter architecture can then be used to predict other co-regulated genes in the genome.

Template methods extrapolate from mapped binding sites in transcriptional control regions

Since a set of transcription factors often regulates multiple target genes, the careful experimental dissection of a single transcriptional control region can be used as a query for discovering other co-regulated genes. These methods initialized a promoter architecture template with the sequences and distance constraints between mapped transcription factor binding sites in promoters of human genes (Klingenhoff *et al.*, 1999; Werner *et al.*, 2003). Genome-wide searches found instances of this template in other proximal promoters. Genes with similar functional annotations to the query sequence were then included in an expanded training set. The template's sequence affinities and distance constraints were adjusted to include all examples within the training set. Although this refined template predicted other co-regulated genes, regrettably none of these predictions were experimentally verified for gene expression similarity.

In a similar approach, a template for neurogenic ectoderm expression in *Drosophila melanogaster* was summarized from four experimentally characterized enhancers (Erives and Levine, 2004). These enhancers shared binding sites for the transcription factors, Twist and Dorsal, as well as an additional sequence motif. The sequence template used a maximal distance of 20 bp between Twist and Dorsal binding sites, as well as a maximal distance of 150 bp between Twist and the novel sequence element. A search of the mosquito (*Anopheles gambiae*) genome found a template match near the ortholog to the *vnd* target gene. This match is striking because sufficient mutations and insertions have accumulated to make the orthologous intergenic regions unalignable. This predicted sequence region from *Anopheles* was sufficient to drive expression in the neurogenic ectoderm of *Drosophila* embryos. Notably, this study suggested that promoter architecture may be conserved even among distant species.

Since sequence templates are typically trained on a handful of genes, they may be prone to computational overfitting. In other words, these template models may fail to capture the full spectrum of transcription factor combinations that can generate a particular gene expression pattern. For instance, a survey of skeletal muscle gene expression identified six transcription factors that were involved in regulation (Wasserman and Fickett, 1998). Since different transcription factor combinations were used at various promoters, no single sequence template could successfully predict all muscle-specific genes. To provide a larger sample of co-expressed genes for model cross-validation, computational methods could analyze sets of co-expressed genes identified by genome-wide expression experiments.

Structured motifs can be discovered from a group of co-expressed genes

For cases where the relevant transcription factors are unknown, computational tools have been developed to search for structured sequence pairs. These tools are extensions of motif discovery methods to find individual transcription factor binding sites. These methods modify their objective functions to search for two sequences simultaneously, often imposing a distance range between them. For instance, the enumerative approach of finding enriched consensus sequences was modified to search for two sequences with a fixed spacing between them (van Helden *et al.*, 1998). Other researchers have exploited a suffix tree data structure to compile putative consensus sequences that are separated by a range of distances (Marsan and Sagot, 2000; Eskin and

Pevzner, 2002). Gibbs sampling methods have also been altered to search for two position weight matrices that are separated by a distance range (Liu *et al.*, 2001; GuhaThakurta and Stormo, 2001). Finally, hidden Markov models can be trained to search for ordered pairs of transcription factor binding sites (Pavlidis *et al.*, 2001). Since these tools focus solely on sequence pattern discovery, they are best used for analyzing groups of co-expressed genes from high-throughput gene expression data. If discovered sequences correspond to known specificities of transcription factors, one could test whether those transcription factors directly regulate the group of co-expressed genes. Whereas these tools use close spacing as a test condition for transcription factors that may bind cooperatively, they simply summarize the range of distances present in the input sequences. Therefore, these methods do not make explicit predictions about distance constraints between binding sites.

Statistical filtering of binding site combinations for similar expression

Evaluating gene expression similarity for target genes that share similar features

All of the computational methods considered so far use a group-by-expression approach. These methods start from a set of co-expressed genes, and then search for enriched sequence features in their transcriptional control regions. Conversely, group-by-sequence approaches can systematically assess the regulatory information associated with an arbitrary sequence feature, such as an individual motif or a pair of sequences. These methods first search for groups of genes that contain a shared sequence feature in their transcriptional control regions. Each gene group can then be statistically evaluated for a higher similarity in gene expression than expected by chance. Significant sequence features are thus predicted to be regulatory elements for that gene group. Note that an advantage of these methods is their ability to discover groups of co-regulated genes.

Transcription factor interactions can be inferred by enumerating pairs of transcription factors and evaluating for gene expression similarity among the predicted target genes. A pioneering study compiled 329 position weight matrices for known or predicted transcription factors and associated a gene group with each possible pair of transcription factors (Pilpel *et al.*, 2001). The expression coherence metric for gene expression similarity was defined as the average Euclidean distance, in the log space of gene expression ratios, between all pairs of genes in the gene group. Transcription factor pairs were inferred to interact if their target genes showed a higher expression coherence value than expected, based on random sampling from the target genes of individual transcription factors. Fifteen examples were discovered for which both transcription factors were known; additional pairs were also reported with only one known transcription factor. Nearly 20% of these pairs showed a statistically significant preference in the order of predicted binding sites. This study also reported only one sequence pair with close spacing preferences: the PAC and RRPE elements found in the promoters of ribosomal RNA transcription and processing genes that are repressed in multiple environmental stress conditions (Hughes *et al.*, 2000a; Gasch *et al.*, 2000).

Predicting gene expression patterns from sequence features

We can assess our understanding of transcriptional regulation by attempting to predict gene expression data from sequence data (Bussemaker *et al.*, 2001; Keles *et al.*, 2002; Conlon *et al.*, 2003). First, all possible sequences of a given length are enumerated. Each sequence can be evaluated for its regulatory potential by correlating its number of occurrences in promoter regions with the corresponding gene's expression change. For instance, sequences that occur more frequently in the promoters of induced genes could be binding sites for activating transcription factors. Conversely, sequences that are enriched in the promoters of repressed genes could be recognized by repressive transcription factors. The slope of the correlation represents the average regulatory contribution of that sequence to gene expression. The most significant word can be added to a regression model that predicts gene expression ratios based on the occurrences of various sequences in promoter regions. After multiple rounds of data fitting, this procedure generates a list of regulatory sequences, their predicted contribution to activation or repression, and the overall fit of the model. Sequences selected in these models generally explain less than 20% of gene expression changes (Conlon *et al.*, 2003). The performance of these models is hampered by random instances of regulatory sequences that are not bound by transcription factors *in vivo*. By using sequence conservation as a filter for instances of regulatory sequences that are non-functional, regression model fits significantly improve (see Chapter 3).

Physical proximity of predicted binding sites may suggest transcription factor interactions

As discussed previously, the close spacing of binding sites may be required for cooperative binding or the efficient recruitment of multiprotein regulatory complexes. Methods that evaluate the proximity of predicted binding sites could potentially identify transcription factor interactions. After predicting all binding site occurrences in a single genome, various statistical tests can evaluate whether the binding sites for any given pair of transcription factors are more closely spaced than expected by chance. For instance, binding sites for Mcm1 or Ste12 alone were found to be randomly distributed in the yeast genome, using an inhomogenous Poisson distribution as a background spacing model (Wagner, 1999). Fourteen promoters contained closely-spaced binding sites, suggesting joint regulation by these two transcription factors. However, only 5 of these 14 genes were shown to be induced in response to alpha factor in a microarray study (Roberts *et al.*, 2000).

A study of human transcription factors in the TRANSFAC database reported 191 transcription factor pairs whose binding sites were enriched for co-localization within a 20-bp window, as assessed by a chi-square test for the occurrence of individual binding sites (Qiu *et al.*, 2002). By comparing their predictions to experimentally verified interactions in the COMPEL database, the authors report a high false negative rate of 62% (Kel-Margoulis *et al.*, 2002b). Another similar study reported 321 pairs that co-localized within a 50-bp or 200-bp window with higher chi-square scores than known interactions in the COMPEL database (Hannenhalli and Levy, 2002). Literature searches on a sample of 100 predicted interactions revealed support for 39 predicted interactions, of which 15 involved protein-protein interactions between the transcription factor pair. These analyses could be improved by optimizing the model parameters, namely the distance between binding sites and the similarity score of predicted binding sites to transcription factor position weight matrices. Nevertheless, these studies demonstrate that the physical proximity of binding sites can be a useful parameter for the discovery of authentic transcription factor interactions.

Distance constraints, along with sequence combinations, were explicitly tested in another sequence-based search for yeast transcription factor interactions (Beer and Tavazoie, 2004). After training 615 position weight matrices from 49 clusters of co-expressed genes, the authors searched for combinations of predicted binding sites that could classify genes back into their 49 expression clusters. While this procedure suffered from circularity and overfitting, it did explicitly test whether several distance constraints improved the accuracy of predictions: the distance of individual sites to translation start in 20 bp windows; the orientation of individual sites; copy number; and distances between pairs of predicted binding sites. Only one example of distance constraints between binding sites was discussed: the PAC element was predicted to occur up to 100 bp upstream of the RRPE element.

Future directions for computational models of multifactorial regulation

I have reviewed computational approaches that explicitly model the effects of multiple transcription factor binding sites on gene expression. Whereas current models can be useful in identifying putative regulatory sequences, they make a couple of key assumptions that fail to incorporate biological knowledge of multifactorial regulation. Since these models assume that regulatory sequences act independently, they fail to account for synergistic interactions between transcription factors. One extension of this method is to evaluate whether pairwise interaction terms between sequences also significantly correlate with gene expression (Keles et al., 2002). These methods also assume that all positions within promoter regions are equivalent. However, distances between binding sites to the translation start site or to other binding sites have been demonstrated to influence gene regulation. Thus, the fidelity of computational models to biological mechanisms would be improved by explicitly modeling distance effects.

Although computational approaches can detect global trends between regulatory sequences and gene expression, they seldom make predictions about specific systems. For example, the regression models described above identified only one pair of regulatory sequences in yeast that may be governed by close spacing. The relatively small sizes of co-regulated gene groups – which typically include fewer than 30 yeast genes – may not provide enough statistical power to detect enriched sequence features. Nevertheless, sequence analyses alone cannot distinguish whether the relevant transcription factors interact by direct protein-protein interactions or by an indirect mechanism, such as collaborative recruitment or nucleosome occlusion. These limitations underscore the need for experiments to test computational predictions.

CHAPTER 2

VISUALIZING ASSOCIATIONS BETWEEN GENOME SEQUENCES AND GENE EXPRESSION DATA USING GENOME-MEAN EXPRESSION PROFILES

PREFACE

Computational analyses of transcription control regions must include a method to evaluate the influence of any given regulatory sequence on gene expression. In this chapter, I describe such a method that assesses the statistical significance of the average gene expression change among genes that share a particular DNA sequence in their upstream regions. As a proof of principle, analyses of previously published gene expression datasets shows how this method can identify sequences that correspond to transcription factor binding sites in yeast.

This chapter appeared in 2001 as an article by myself, Pat Brown and Mike Eisen in the journal *Bioinformatics*, volume 17, pages S49-S55. Further GMEP analyses of the Rosetta compendium microarray data are reported for the first time in Table 2.2.

ABSTRACT

The combination of genome-wide expression patterns and full genome sequences offers a great opportunity to further our understanding of the mechanisms and logic of transcriptional regulation. Many methods have been described that identify sequence motifs enriched in transcription control regions of genes that share similar gene expression patterns. Here we present an alternative approach that evaluates the transcriptional information contained by specific sequence motifs by computing for each motif the mean expression profile of all genes that contain the motif in their transcription control regions. These genome mean expression profiles (GMEP’s) are valuable for visualizing the relationship between genome sequences and gene expression data, and for characterizing the transcriptional importance of specific sequence motifs. Analyses of GMEP’s calculated from a dataset of 519 whole-genome microarray experiments in *Saccharomyces cerevisiae* show a significant correlation between GMEP’s of motifs that are reverse complements, a result that supports the relationship between GMEP’s and transcriptional regulation. Hierarchical clustering of GMEP’s identifies clusters of motifs that correspond to binding sites of well-characterized transcription factors. The GMEP’s of these clustered motifs have patterns of variation across conditions that reflect the known activities of these transcription factors.

BACKGROUND

As genome sequencing projects move forward at a rapid pace, and as the use of DNA microarrays and related techniques becomes more widespread, there are a growing number of organisms for which both complete genome sequences and large volumes of genome-wide transcript abundance measurements are available. An obvious challenge in the analysis of these data is to understand the cellular mechanisms used to orchestrate genomic expression programs. As complex models of transcriptional networks have yet to reach maturity, most recent research has focused on the more modest goal of using genome-wide expression patterns and genome sequences to identify likely (and ideally previously unidentified) transcription factor binding sites.

Most common strategies adopt a “group-by-expression” approach, in which genes with similar expression are identified, and then their transcription control regions are analyzed for the presence of shared sequence motifs (reviewed in Ohler and Niemann, 2001). These approaches postulate that genes with similar patterns of expression are likely to be regulated by common factors, and thus should share binding sites for these factors in their non-coding regions. Co-expressed genes are identified by cluster analysis of gene expression data (c.f. DeRisi *et al.*, 1997; Spellman *et al.*, 1998; Cho *et al.*, 1998; Tavazoie *et al.*, 1999; Gasch *et al.*, 2000). Sequences upstream of co-expressed genes are analyzed for statistically over-represented sequence motifs using a variety of algorithms, including: expectation maximization (Bailey and Elkan, 1995), over-represented oligomers (van Helden *et al.*, 1998; Wolfsberg *et al.*, 1999), weight matrices (Hertz and Stormo, 1999), Gibbs sampling (Hughes *et al.*, 2000a), enumerative statistics (Sinha and Tompa, 2000), probabilistic segmentation (Bussemaker *et al.*, 2000), and sequence pattern discovery (Vilo *et al.*, 2000). As has as been previously noted (Holmes and Bruno, 2000; Wagner, 1999), a problem with this approach is that it does not take into account the multiple, independent mechanisms by which most genes are regulated. For example, two genes can be co-regulated under one set of conditions, but differentially regulated under others. Although these genes would not be easily identified as co-expressed, they nonetheless share important regulatory information.

An alternate strategy is to adopt a “group-by-sequence” approach in which the transcriptional control content of sequence motifs is evaluated on the basis of the expression patterns of genes that contain the motif in their nominal transcription control regions (TCR; the adjacent cis-DNA that is believed to contain sequences that determine the transcriptional regulation of the gene). If a sequence motif carries transcriptional information – namely if it is bound by a transcription factor and this binding alters the transcription rate of adjacent genes – we expect the expression patterns of genes that contain this motif in their TCR’s to have non-random features that reflect the activity of the corresponding factor. In contrast, if a motif does not encode transcriptional regulatory information, the genes that contain the motif in their TCR’s should not have expression patterns that differ significantly from those of the entire population of genes.

To evaluate and exploit this expectation, we define the genome mean expression profile (GMEP) of a sequence motif as the mean expression profile of all genes (regardless of the expression profiles) that contain this motif in their TCR’s. To understand the reasons for using GMEP’s, consider a set of genes whose transcription is increased by the activity of a given transcription factor in some set of conditions, but whose expression patterns are otherwise unrelated. Although the multi-factorial nature of transcription control could easily obscure the commonalities in these genes' expression profiles, we nonetheless expect that, on average, these genes will have higher expression levels in the conditions where this transcriptional activator is active when compared to some randomly chosen subsets of genes, and we expect the magnitude of this elevation will reflect the activity level of the activator. Additional genes that contain this motif in their TCR's but which are not regulated by the particular activator should also have mean expression profiles that are close to the population mean profile. Thus, the GMEP of the sequence motif recognized by the activator should differ significantly from the population mean profile only when the activator is present and active and this difference should be greatest when the activator has its highest level of activity. Note that this should still be true even if transcription of the regulated genes is also independently and separately controlled by additional non-overlapping factors.

ALGORITHM

Data

To compute GMEP's, we begin with a data matrix D with r rows, each representing a single gene, and c columns, each representing a single experimental condition. Each cell D_{gj} represents the expression level of gene g in condition j. Missing values are allowed. In the data used here these values are log-transformed (base 2) relative expression ratios (compared to a suitable reference sample) and the columns are mean-centered. For each gene, we define a sequence $S(g)$ that is the genome sequence of the gene's nominal transcription control region. Note that for most organisms there are no well-defined rules for identifying TCR's; for analyses presented here using the yeast *Saccharomyces cerevisiae,* $S(g)$ is the 600 basepairs upstream of the translation start site for gene g .

Genome-mean expression profiles

For a DNA sequence motif m, let G be the set of genes that contain this motif in their TCR's. Define the **genome-mean expression profile** of motif m [denote GMEP(m)] to be the c-dimensional vector equal to the weighted mean of the c-dimensional vectors that represent the expression profiles of each gene in *G*:

$$GMEP(m)_j = \frac{\sum_{g}^{g\in G} w_{mg} \times D_{gj}}{\sum_{g}^{g\in G} w_{mg}}$$

where w_{mg} is the number of occurrences of motif m in $S(g)$. A weighted mean was used since transcription factors may have a higher affinity to genes that contain multiple copies of their cognate sites (Wagner, 1999).

For simplicity, here we only enumerate motifs containing the symbols A, C, G, or T although this is not a necessary constraint. For a given data matrix D and a fixed motif length L, we compute the ($4^L \times c$) matrix where each row is the GMEP of a single motif. To correspond with the data matrix D, the columns in the GMEP matrix are mean-centered.

Significance testing

To analyze the likelihood that specific values in our GMEP matrixes are expected to have occurred by chance we compute approximate Z-scores for hypothesis testing. Consider the calculation of a GMEP as the mean of a sample ($X_1, \ldots, X_n$) of n gene expression levels drawn randomly (with replacement) from a population. This population comprises all relative gene expression measurements from a single microarray experiment. If a motif does not contain transcriptional information, the expression levels of genes that contain this motif in their TCR’s represent a randomly drawn sample, and the GMEP for this motif should not differ significantly from the population mean. Alternatively, if a motif does contain transcriptional information, and the corresponding transcription factor(s) are active, then we expect the GMEP to be different from the population mean.

Assume that $X_1, \ldots, X_n$ are independent and identically distributed. We expect that the sample means of many samples chosen randomly from a given (microarray) population with mean μ and standard deviation s will fall on a normal distribution with mean μ and standard deviation $\frac{s}{\sqrt{n}}$. Using the above sampling distribution as a null distribution, we can approximate a *Z*-score for each value in the GMEP data matrix:

$$Z_{m\,j} = \frac{GMEP(\,m\,)_j - \mu_j}{s_j \,/\, \sqrt{n_m}}$$

where n_m is the number of observations for motif m that went into the mean, and μ_j and s_j are the mean and standard deviation, respectively, of all relative expression measurements for microarray experiment j.

However the above assumption may not be valid because: (1) relative gene expression levels for individual genes may be correlated; (2) the distribution of w_{mg} is not uniform across the genome. In this case, the mean GMEP value is still expected to equal μ, but the standard deviation of the GMEP value will vary. We investigated the variability of the standard deviation from the predicted value of $\frac{s}{\sqrt{n}}$ by permutation tests. For each of 51 hexameric motifs, a distribution of randomized GMEP values was obtained from 5000 random permutations of the d_{gj} values for a single experiment. Since the mean and standard deviations for these permutation distributions varied by less than 5% from the values given by the central limit theorem, we assume that the above equation provides a good approximation to the *Z*-score.

RESULTS

We computed GMEP's for all motifs of length 5, 6, 7, or 8 nucleotides in length using an input gene expression dataset of 517 different DNA microarrays, each containing ~5300 yeast genes (overlapping genes and duplicated genes were not included in this analysis). This dataset comes from the Stanford Microarray Database (Sherlock *et al.*, 2001) and includes the published results of DeRisi *et al.*, 1997; Spellman *et al.*, 1998; Chu *et al.*, 1998; Gasch *et al.*, 2000; Ogawa *et al.*, 2000 and some unpublished results that will be described in a forthcoming publication.

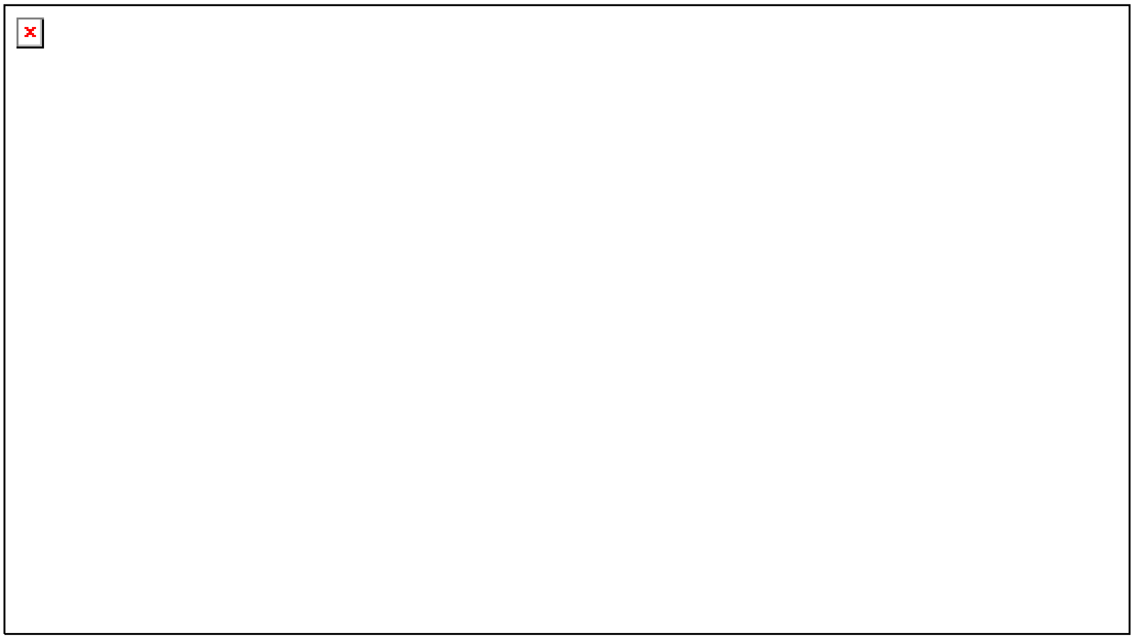
Many asymmetric transcription factor binding sites confer similar regulation irrespective of their orientation relative to the target gene. If GMEP's reflect transcriptional regulation associated with a sequence motif, then we would expect GMEP's for many *bona fide* regulatory motifs to be highly correlated with the GMEP's of their reverse complements (note that in computing GMEP's we only use motifs found on the positive strand of adjacent non-coding DNA, so there is no *a priori* expectation that GMEP's of reverse complementary motifs should be correlated). We computed the correlation between all motifs and their reverse complements (excluding motifs that are self-reverse complements) using the GMEP matrixes described above, and compared the results to a negative control in which the associations between cis-sequences and gene expression vectors were randomly permuted. Since this control maintained the same set of expression profiles and regulatory sequences (only the assignment of the expression profiles to each motif were permuted) effects due either to the expression patterns themselves or to the distribution of motifs in non-coding sequences would be found in both the real and permuted data.

Figure 2.1 shows the histograms of reverse-complement correlations for motifs of lengths 5, 6, 7, and 8. As expected, for all of these motif lengths, the distributions of reverse-complement correlations for the randomly permuted datasets resembled normal distributions with mean values close to zero. In contrast, when the correct associations of gene expression data were used, there was a striking shift in the distribution of complement correlations towards positive correlations, with a distribution mean ranging from 0.357 (motif length 5) to 0.080 (motif length 8). The positive correlations in the GMEP's between many of the motifs and their reverse complements support the assumption that many regulatory motifs encode information when present on either of the DNA strands and validates the biological relevance of GMEP's.

The distributions of reverse-complement correlations displayed motif-position dependence. We computed the distribution of reverse-complement correlation values for hexameric motifs found in 100 bp windows between –1000 and +1000 bp relative to the translation start site. Figure 2.2 shows the reverse-complement correlations associated with motifs found at different positions. The highest reverse-complement correlations occur for motifs found between –100 and –200 of the translation start site, while the reverse-complement correlation decays to near-background as the distance from the start site increases. This result agrees with other data on the positional distribution of transcriptionally active transcription factor binding sites in yeast (Wolfsberg *et al.*, 1999).

We chose to examine in more detail the data for all 4096 possible hexameric motifs. After calculating the GMEP associated with each motif, we then organized this data using hierarchical clustering (Eisen *et al.*, 1998). The logic of applying clustering to GMEP's was that motifs that encode similar regulatory information would display similar

Figure 2.1) Distributions of complement correlations for all motif / reverse complement pairs



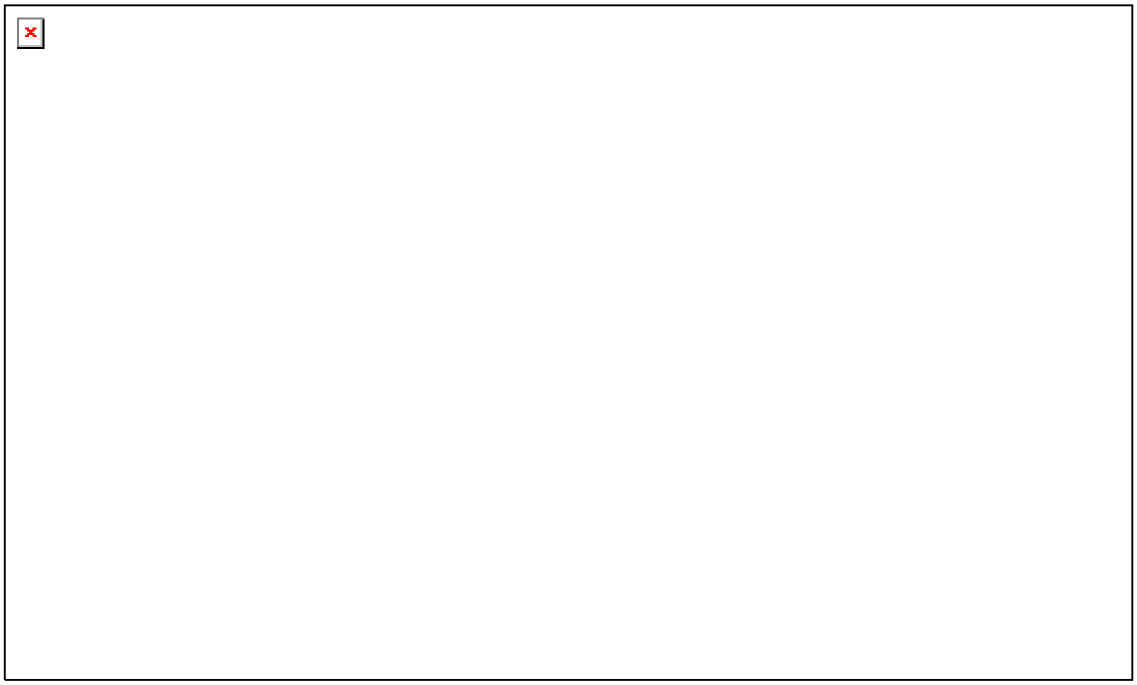
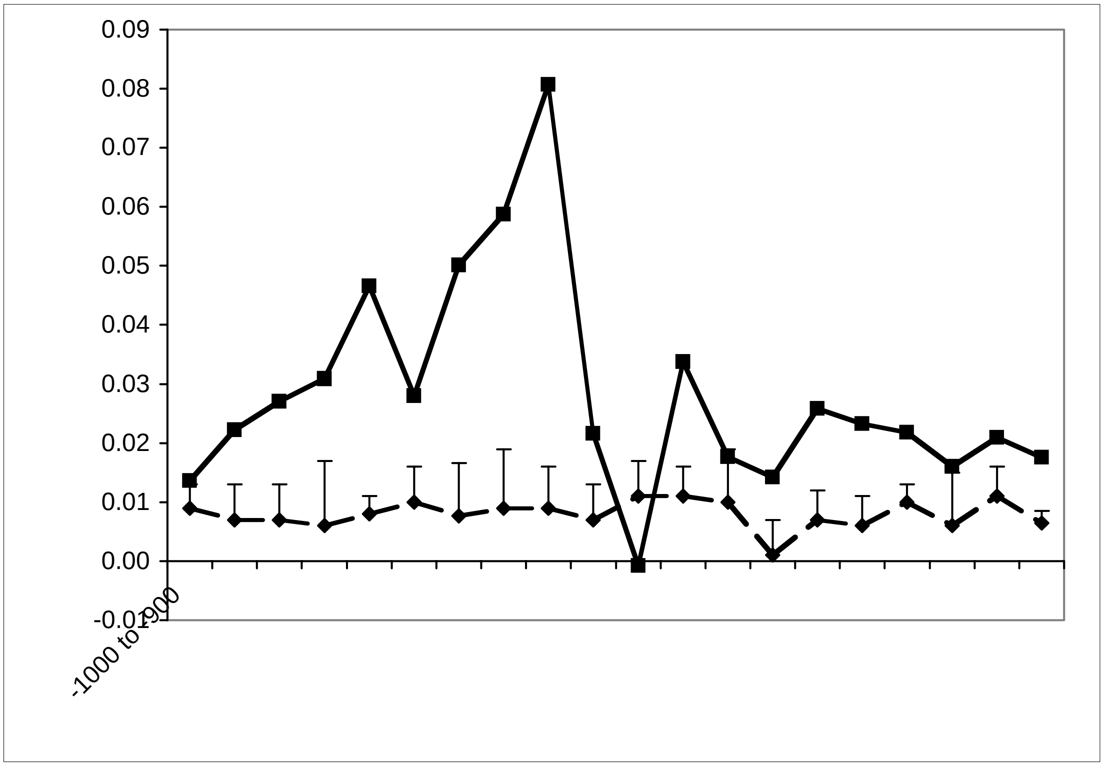


Figure 2.1 (continued)

Correlations between the GMEP for a motif and the GMEP of its reverse complement were calculated as described. Dashed lines indicate the distribution of Pearson coefficients for randomly permuted associations between each non-coding region and a gene expression profile, whereas solid lines indicate the distributions for actual data. Mean values of the actual vs. randomized distributions: 0.357 vs. 0.009 (length 5); 0.245 vs. -0.006 (length 6); 0.148 vs. 0.003 (length 7); 0.080 vs. -0.001 (length 8).

Figure 2.2) Position-dependent effects of shifts in the complement correlation distribution



Mean values were calculated for binding signal distributions in 100 bp windows at varying distances away from the translation start site. Open lines with diamonds represent the mean of mean values obtained from five different trials in which the association between each non-coding sequence and gene expression level was randomly permuted. The error bars indicate the standard deviation for these five trials. Filled lines with squares represent the mean values of the binding signal distribution for actual data.

GMEP’s and would thus be clustered together, and that motifs within a cluster might comprise different submotifs of a single consensus binding site.

We found that examining the clustered GMEP data in TreeView (Eisen *et al.*, 1998), provided an efficient way to visually identify clusters of motifs associated with biologically interesting expression patterns. We used two stringent heuristic criteria for identifying “interesting” motif clusters: (1) the GMEP’s within the clusters had correlations with each other of greater than 0.75; and (2) the motifs within each cluster were orientation-independent (*i.e.* each cluster containing at least one reverse complement pair with a correlation greater than 0.7). Table 2.1 lists nine separate motif clusters that met these criteria. Each of these clusters contains previously-identified promoter motifs, including the MCB element bound by the MBF transcription factors, the STRE element recognized by the Msn2 and Msn4 transcription factors, and a site involved in environment stress response that has been previously identified but whose putative binding factor remains unknown. Figure 2.3 shows the GMEP clusters associated with these motifs. These GMEP profiles reflect conditions in which these transcription factors are known or believed to be active. GMEP analysis on another dataset of gene expression for over 300 mutants of *Saccharomyces cerevisiae* discovered additional transcription factor binding sites (Table 2.2) (Hughes *et al.*, 2000b).

Table 2.1) Examples of sequences that regulate gene expression in many conditions

Consensus Sequence
R = [A/G];
S = [C/G];
W = [A/T]
	Sequence motif	Cluster correlation	Mean complement correlation	Characteristics of GMEP

TGAAAATTTT	RRPE[1]	0.974	0.977 (*n*=4)	Generally repressed
AWTTTTCWTTT	RRPE[1]	0.963	0.809 (*n*=14)	Generally repressed
SCACGTG	Pho4[2]	0.775	0.628 (*n*=6)	Induced in Δ*pho80*, Δ*pho85* mutants
TGASTCA	Gcn4[2]	0.751	0.698 (*n*=3)	Induced during amino acid starvation
AGGGG	Msn2 or Msn4	0.897	0.880 (*n*=26)	Induced during stress
ARGGGAWA	Msn2 or Msn4	0.840	0.794 (*n*=15)	Induced during stress
CAG[C/A]GATG AG[C/A]T	Unknown[3]	0.834	0.880 (*n*=20)	Repressed during stress
WCGCGW	Mbp1-Swi4[4]	0.814	0.757 (*n*=11)	Cell cycle periodicity
GATAAG	Gln3[2]	0.810	0.854 (*n*=2)	Induced during amino acid starvation

[1] Hughes *et al.*, 2000a
[2] van Helden *et al.*, 1998
[3] Gasch *et al.*, 2000
[4] Spellman *et al.*, 1998

Consensus sequences were assembled from the individual motifs comprising clusters that were selected using the criteria described in the text. Cluster correlation refers to the Pearson correlation among corrected GMEP's for all motifs found in the cluster. The mean binding signal refers to the mean value for the Pearson correlation between the uncorrected GMEP of each motif found in the cluster with the uncorrected GMEP of its reverse complement. The number of motifs in each cluster is indicated in parentheses.

Table 2.2) Examples of sequences that regulate gene expression in the Rosetta compendium dataset

Consensus Sequence	Transcription factor	Selected Mutant strains with significant GMEP values
WTGCTGG	Ace2 or Swi5	(-): ERG11^{C}
TGCACCCG	Aft1 or Rcs1	(+): *vma8*, *cup5*, *mac1*, *rip1*
CCTCGAGG	DRC	(+): *dig1 dig2*, *gas1*, *spf1*, *anp1*, *swi4*, *she4*
TGASTCA	Gcn4	(+): ERG11^{C}, *erg2*, *erg3*
WCGCGW	Mbp1 or Swi4	(-): ERG11^{C}, *erg2*, *erg3*, itraconazole
CCCCGC	Mig1	(+): *tup1 ssn6*, *ycr050c*, *ymr273c*, *acp2*, *kin4*, *vac8*, *bim1*
TCCGCGGA	Pdr1 or Pdr3	(+): *bub3*, *cem1*, *afg3*, *rml12*, *aep2*, *kim4*
SCACGTG	Pho4	Many conditions
GCACCC	Rap1	(+): *cem1*, *afg3*, *rml2*, *aep2*, *kim4*, *imp2*, *pet111*, *cyt1* (-): *top1*, *yel033w*, *dot4*, *rpl34a*
GGTCACG	Rtg1 or Rtg3	(-): *rtg1*
TGAAACA	Ste12	(+) *dig1* (-): *ste5*, *ste11*, *ste18*, *ste7*, *ste12*, *ste24*
CCTCGTA	Upc2	(+): ERG11^{C}, HMG2^{C}, lovastatin, itraconazole, terbinafine

IUPAC symbols: S = C or G; W = A or T

Figure 2.3) Examples of GMEP clusters

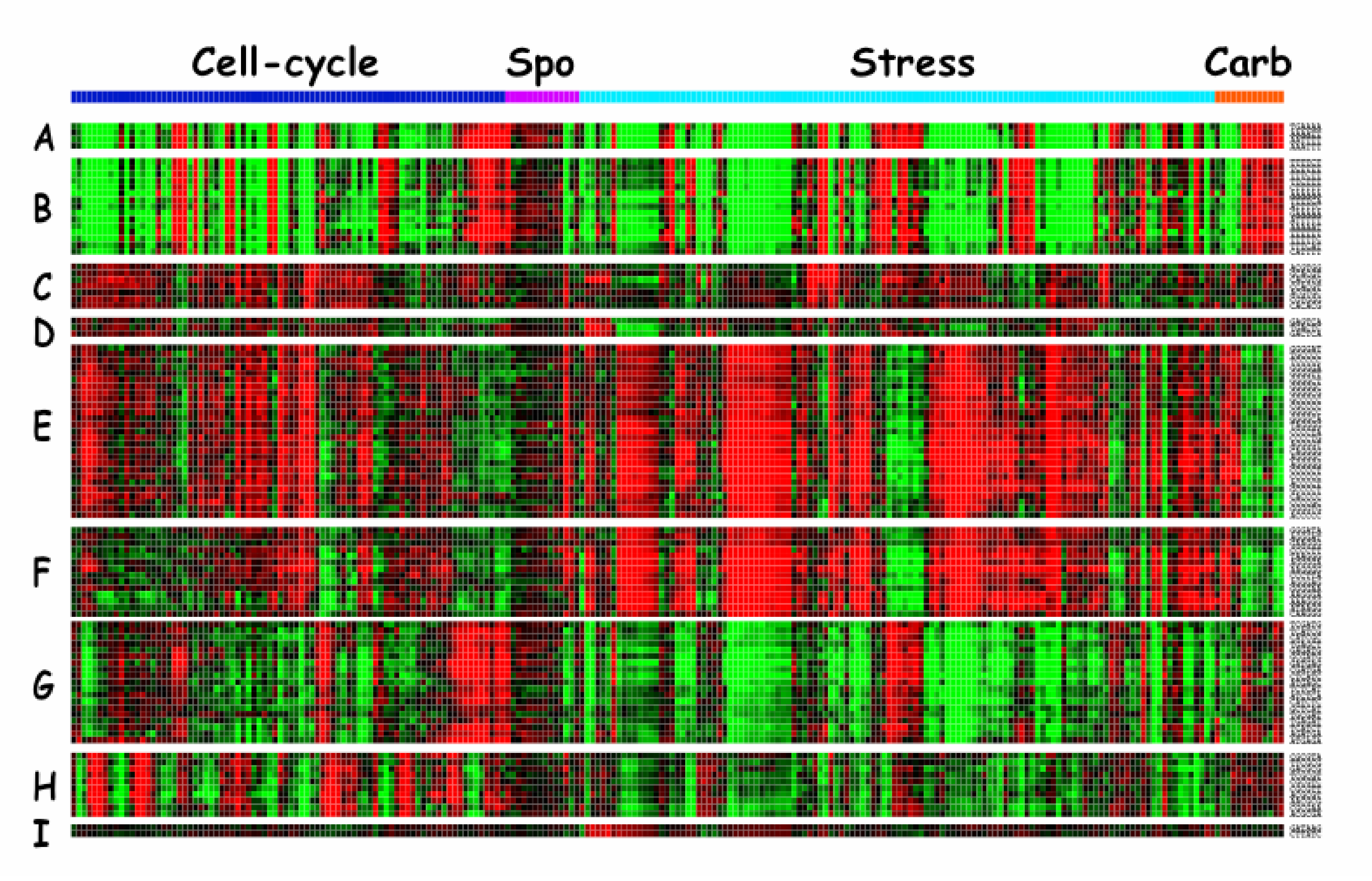


Figure 2.3 (continued)

Each row represents the GMEP for a single motif, calculated using equation (1). Each column represents a single cDNA microarray experiment. The columns selected for display correspond to microarray experiments for the mitotic cell cycle conditions (Cell-cycle: Spellman *et al.*, 1998; Zhu *et al.*, 2000); sporulation conditions (Spo: Chu *et al.*, 1998); environmental stress conditions (Stress: Gasch *et al.*, 2000; Ogawa *et al.*, 2000); and alternate carbon sources (Carb: Spellman PT, Brown PO and Botstein D, unpublished observations). Green pixels correspond to transcriptional repression, red pixels correspond to transcriptional induction, and the pixel intensity reflects the magnitude of the change in average gene expression.

Nine clusters of GMEP's are displayed that meet our selection criteria as discussed in the text. Manual alignment of these motifs yields the following consensus sequences, which are also listed in Table 1: (A) TGAAAATTTT (RRPE); (B) AWTTTTCWTTT (RRPE-like); (C) SCACGTG (Pho4); (D) TGASTCA (Gcn4); (E) AGGGG (Msn2 or Msn4); (F) ARGGGGAWA (Msn2- or Msn4-like); (G) CAG[C/A]GATGAG[C/A]T (Repressed in stress – Gasch *et al.*, 2000); (H) WCGCGW (Mbp1-Swi4); (I) GATAAG (Gln3).

DISCUSSION

Genome mean expression profiles represent one of several alternatives to "group-by-expression" approaches for analyzing gene expression data. Rather than look for statistical over-representation of sequences in a fixed subset of genes, these alternative methods introduce conceptual models that underlie microarray data. Holmes and Bruno (2000) have developed a likelihood framework to consider similarities in both sequences and gene expression profiles at the same time. The clustering of genes can thus by guided by choosing the most likely sequence-expression model that yields the observed gene sequences and gene expression levels. Bussemaker *et al*. use a regression method to fit gene expression data to a multivariate linear model. Significant motifs are defined to be those that yield the largest reduction in the χ^2-squared statistic.

The genome mean expression profile introduced here is a simple and straightforward tool for assessing the information content of sequence motifs. The underlying model is a simple one. However, the observed correlation between reverse-complement pairs, the striking position-dependence of this correlation, and the success in identifying many known transcription factor binding sites strongly support continued analysis of the current data and the development of more sophisticated derivatives.

We identified more words that matched known transcription factor binding sites in a gene expression dataset of deletion strains grown in rich media, compared with gene expression of wild-type cells in multiple stress conditions (Hughes *et al.*, 2000b). Whereas our method averages out effects of multifactorial regulation, it performs better when analyzing the transcriptional response to targeted genetic ablations. In the Rosetta compendium dataset, the identified binding sites often corresponded to gene expression changes in separate signaling pathways from the primary mutation. For instance, genes regulated by the iron-regulated transcription factor Aft1 were induced on average in deletion mutants of the copper-regulated transcription factor (*mac1Δ*). Deletion mutants of the ergosterol biosynthesis pathway showed repression of genes regulated by the cell cycle (Mbp1/Swi4 targets) or by general amino acid control (Gcn4 targets). These transcriptional responses suggest that cross-regulation occurs among signaling pathways during compensation for genetic mutants.

CHAPTER 3

PHYLOGENETICALLY AND SPATIALLY CONSERVED WORD PAIRS ASSOCIATED WITH GENE EXPRESSION CHANGES IN YEASTS

PREFACE

This chapter reports a computational screen for transcription factor pairs that participate in multifactorial regulation. Motivated by the expectation that promoter architecture has been phylogenetically conserved, I developed sequential statistical tests to find conserved word pair templates with co-conservation and close spacing of DNA sequence pairs. By extending the group-by-sequence approach from the previous chapter to associate these templates with significant gene expression changes, I identified several examples of multifactorial regulation in yeasts.

This chapter appeared in 2003 as an article by myself, Alan Moses, Manolis Kellis, Eric Lander and Mike Eisen in the journal *Genome Biology*, volume 4, article R43.

ABSTRACT

Background

Transcriptional regulation in eukaryotes is often multifactorial, involving multiple transcription factors binding to the same transcription control region (*e.g.*, upstream activating sequences and enhancers), and to understand the regulatory content of eukaryotic genomes it is necessary to consider the co-occurrence and spatial relationships of individual binding sites. The identification of sequences conserved among related species (often known as phylogenetic footprinting) has been successfully used to identify individual transcription factor binding sites. Here, we extend this concept of functional conservation to higher-order features of transcription control regions involved in the multifactorial control of gene expression.

Results

We used the genome sequences of four yeast species of the genus *Saccharomyces* to identify sequences potentially involved in multifactorial control of gene expression. We found 989 potential regulatory “templates”: pairs of hexameric sequences that are jointly conserved in transcription regulatory regions and also exhibit non-random relative spacing. Many of the individual sequences in these templates correspond to known transcription factor binding sites, and the sets of genes containing a particular template in their transcription control regions tend to be differentially expressed in conditions where the corresponding transcription factors are known to be active. Several templates correspond to pairs of transcription factors known to act together, while others suggest previously uncharacterized pairs of transcription factors that may work coordinately. The incorporation of word pairs to define sequence features yields more specific predictions

of average expression profiles and more informative regression models for genome-wide expression data than considering sequence conservation alone.

Conclusions

The incorporation of both joint conservation and spacing constraints of sequence pairs predicts groups of target genes that were specific for common patterns of gene expression. Our work suggests that positional information, especially the relative spacing between transcription factor binding sites, may represent a common organizing principle of transcription control regions.

BACKGROUND

All organisms have evolved intricate signaling networks that sense and respond to their environment. At a cellular level, the activation of one or more signaling networks often leads to coordinated changes in gene expression, via the regulated activity and binding of transcription factors to transcription control regions (TCRs) of genes (*e.g.* enhancers and upstream activating sequences). In yeast and most other eukaryotes, the transcriptional regulation of individual genes is often multifactorial, as multiple transcription factors may bind to a single TCR (reviewed in Wolberger, 1999). Multifactorial regulation encompasses several distinct biochemical mechanisms. In some cases, transcription factors may bind cooperatively to adjacent DNA sites via direct physical interactions (Bhoite *et al.*, 2002; Mead *et al.*, 2002). In other examples, multiple transcription factors that bind independently may recruit a common co-activator (Blaiseau and Thomas, 1998), or may act independently of one another to alter gene expression in response to distinct cellular cues (Gasch, 2003). Recent studies have also suggested that nearby transcription factors may collaboratively compete with nucleosomes, thus enhancing the binding of individual transcription factors (Vashee *et al.*, 1998; Miller and Widom, 2003). Many experiments in yeast have shown that specific pairs of factors must be bound near each other for multifactorial regulation to occur (Smith and Johnson, 1992; Vashee *et al.*, 1998; Miller and Widom, 2003), and it is on these spatial constraints that we focus here.

The challenges in understanding how regulatory information is encoded in genomes include both the identification of regulatory sequences in TCRs and the elucidation of sequence constraints on productive multifactorial regulation. Previous computational work has been devoted to identifying putative transcription factor binding sites. A plethora of computational methods has been developed to find over-represented sequences in a subset of genes believed to contain a common transcription factor binding site (reviewed in Stormo, 2000). The rapid pace of genome sequencing has enabled a complementary approach – phylogenetic footprinting (reviewed in Duret and Bucher, 1997; Pennacchio and Rubin, 2001) – that recognizes that the conservation of sequences across related organisms often reflects evolutionary selection for their presence in TCRs. Several algorithms have been developed to perform phylogenetic footprinting analyses systematically (Blanchette and Tompa, 2002; Loots *et al.*, 2002; Schwartz *et al.*, 2000).

After compiling a collection of putative binding sites, associations can be made between various binding site assortments and gene expression. Some recent approaches include Boolean logic (Pilpel *et al.*, 2001), regression methods (Bussemaker *et al.*, 2001; Keles *et al.*, 2002; Wang *et al.*, 2002; Conlon *et al.*, 2003), spatial clustering (Wagner, 1999), and multiple binding site matrix classifiers (Klingenhoff *et al.*, 1999; Pavlidis *et al.*, 2001; Kel-Margoulis *et al.*, 2002a). Spatial information on the relative locations of binding sites is ignored in all but the last two classes of approaches. Yet even these methods, which often search for fixed arrangements among the individual binding sites, may miss permutations in the ordering of binding sites within TCRs that may still be bound and regulated by their corresponding transcription factors.

The primary aim of this work was to incorporate positional information and phylogenetic footprinting to identify sequence motifs that may regulate gene expression. Consequently, we expanded the focus of phylogenetic footprinting from the conservation of contiguous sequences to higher-order features of TCRs, namely the spatial organization of individual binding sites. Since transcription factors participating in multifactorial regulation may require physical proximity among their binding sites, we searched for groups of conserved sequences that were more closely spaced in TCRs than expected. We refer to these spatially organized sequences as conserved word templates.

As a proof of principle, we started with the simplest example of such templates: pairs of conserved 6-bp words. Conservation was assessed using the genome sequences of three additional *Saccharomyces* species, which were chosen to be sequenced in order to elucidate regulatory sequences conserved among these closely related species (Kellis *et al.*, 2003). To exploit this comparative genome data, we have devised a method that systematically tested sequence pairs for joint conservation across genomes and close spacing within individual TCRs. Since genes regulated by the same set of transcription factors often displayed similar gene expression patterns in certain experimental conditions, we identified conserved word pair templates whose gene targets were associated with common changes in gene expression. We adopted a group-by-sequence approach to first identify genes that contained the word pair templates and then to test for significant associations with expression levels of the identified genes (Chiang *et al.*, 2001). Significant associations between conserved word pair templates and specific gene expression changes and the prevalence of known transcription factor binding sites suggest that conserved word pair templates comprise sequences important for multifactorial regulation in yeast. In addition, conserved word pair templates represent more specific predictors of gene expression than individual words or word pairs in *S. cerevisiae*.

MATERIALS AND METHODS

Datasets

Whole-genome shotgun sequencing of *Saccharomyces bayanus*, *Saccharomyces mikatae*, and *Saccharomyces paradoxus* has been previously described (Kellis *et al.*, 2003). All of these organisms are highly related to *Saccharomyces cerevisiae*, as they are grouped within the *sensu stricto* branch of the *Saccharomyces* genus (Cliften *et al.*, 2001). Intergenic regions were aligned using CLUSTALW as described (Kellis *et al.*, 2003) and are available from the Saccharomyces Genome Database. A total of 4101 CLUSTALW alignments were analyzed. These alignments were filtered for orthologs in at least 3 genomes.

Gene expression measurements were obtained from the Stanford Microarray Database (Sherlock *et al.*, 2001) and Rosetta (Hughes *et al.*, 2000b). The main experimental types among the 342 conditions examined include cell cycle (Spellman *et al.*, 1998; Cho *et al.*, 1998), environmental stress response (Gasch *et al.*, 2000), DNA damage (Gasch *et al.*, 2001; Lee *et al.*, 2000), cadmium (N. Ogawa and P. O. Brown, unpublished data), and inhibition of ergosterol biosynthesis (Hughes *et al.*, 2000b). This data has been log-transformed (base 2), and each experimental condition has been median normalized.

Dependent conservation of word pairs

To assess whether two words were co-conserved in the same intergenic regions, a chi-square test of independence was systematically conducted for all possible words of length six. We defined a word to include a 6-bp sequence and its reverse complement.

Each transcriptional control region (TCR) for a gene was defined as the 600 base pairs upstream of its translation start site. TCRs shared between divergently transcribed genes less than 600 bp long were only counted once. A word was labeled conserved in a TCR if all six bases were identical among at least three of the four genomes in the CLUSTALW alignment. For each word pair (*W*, *V*) whose overlap was less than 4, a contingency table $\boldsymbol{C}_{wv}$ was constructed. In this table, $\boldsymbol{C}_{wv}$ = # TCR($\boldsymbol{I_w} \cap \boldsymbol{I_v}$), where $\boldsymbol{I_w}$, $\boldsymbol{I_v}$ are indicator variables for the presence of each conserved word in a TCR. TCRs shared between divergently transcribed genes less than 600 bp long were only counted once. The expected counts $\boldsymbol{E}_{wv}$ were obtained from an independence assumption, *i.e.* the product of the individual word conservation probabilities, multiplied by the total number of TCRs. The chi-square statistic with Yates continuity correction was computed according to the definition:

$$\chi^2_{wv} = \sum_{I_w=0}^{1} \sum_{I_v=0}^{1} \frac{\left(\left|C_{wv} - E_{wv}\right| - \frac{1}{2}\right)^2}{E_{wv}}$$

Spatial proximity of constrained word pairs

The second requirement for a conserved sequence template involved constraints on spatial arrangements between individual words. Any method that evaluates spacing distributions between word pairs must take into account positional biases that may be present for individual words (A. M. Moses, unpublished results). We used a permutation test to evaluate the significance of the median of minimum distances, excluding overlaps, between conserved word pairs. By permuting the TCR labels for one of the words, but not the word positions themselves, we retained the positional biases of individual words within intergenic regions. Within any given TCR $\boldsymbol{t}$, define $\boldsymbol{p_t}(\boldsymbol{W}) = \{\boldsymbol{p_t}^{1}(\boldsymbol{W}), \ldots, \boldsymbol{p_t}^{j}(\boldsymbol{W})\}$ as a vector of positions in *S. cerevisiae* where word W is conserved. Suppose that words $\boldsymbol{W}$ and $\boldsymbol{V}$ were jointly conserved in TCRs $\boldsymbol{T_1} \ldots \boldsymbol{T_N}$. For each TCR $\boldsymbol{t} \in \{ \boldsymbol{T_1} \ldots \boldsymbol{T_N} \}$, the minimum distance between words $\boldsymbol{W}$ and $\boldsymbol{V}$ was computed as:

$$m_t = \min_{j,k} \left| p_t^j(W) - p_t^k(V) \right|$$

The median of minimum distances, $\overline{D}$, was simply the median of the ordered distribution $\{ \boldsymbol{m_1}, \ldots, \boldsymbol{m_t} \}$.

We used a permutation test to generate an empirical null distribution of $\overline{D}$ for all word pairs with $N \geq 10$. After randomly permuting the labels $\boldsymbol{t}$ for the position vectors of word V, a permutation test statistic, $\overline{D}^{\boldsymbol{*}}$, can be calculated as above. By repeating this resampling procedure R times, an empirical null distribution $\overline{D}_{null} = \{ \overline{D}^{*1}, \ldots, \overline{D}^{*R} \}$ can be obtained. The significance of the observed median of minimum distances, $\overline{D}$, in the N promoters was calculated as its quantile in the empirical null distribution $\overline{D}_{null}$. We set an upper bound of $R = 10^6$, but stopped permutations early if 20 or more values in $\overline{D}_{null}$ were found less than $\overline{D}$.

Correction for multiple testing involved control of the proportion of false positives using a False Discovery Rate method (Benjamini and Hochberg, 1995). This method has increased power over Bonferroni-type methods. Permutation quantiles for all N word pairs tested were sorted in non-decreasing order: $\boldsymbol{q_1} \leq \ldots \leq \boldsymbol{q_N}$.

Let $k = \max\left(i : q_i < \frac{0.05 \times i}{N} \right)$. Then the first $\boldsymbol{k}$ word pairs in the ordering had a corrected significance level of $q < 0.05$, *i.e.* the rate of false positives is approximately 5%.

Association between template-specified gene groups and gene expression changes

For each gene expression condition $\boldsymbol{c}$ in our dataset, $\boldsymbol{c} \in \{1, \ldots, 342\}$, we tested the null hypothesis that a gene subset $\boldsymbol{G}_{\mathrm{wv}} \subseteq \boldsymbol{G}$ selected by a conserved word pair (w, v) had the same distribution of gene expression ratios ($\boldsymbol{E}_{\mathrm{wv}}{}^{\mathrm{c}}$) as the entire genome ($\boldsymbol{E}^{\boldsymbol{c}}$). The alternate hypothesis stated that the two gene expression distributions were significantly different. Any gene was an element of $\boldsymbol{G}_{\mathrm{o}}$ if its corresponding TCR conserved both sequences in the word pair. Since the size N_{o} of gene subsets may be small and the distributions may not be normally distributed, we used the nonparametric Kolmogorov-Smirnov (K-S) test. The test statistic $\boldsymbol{K}$ compares the cumulative distribution functions $\boldsymbol{F}_{\mathrm{wv}}{}^{\boldsymbol{c}}$ and $\boldsymbol{F}^{\boldsymbol{c}}$ corresponding to $\boldsymbol{E}_{\mathrm{wv}}{}^{c}$ and $\boldsymbol{E}^{\boldsymbol{c}}$ by the formula $\boldsymbol{K} = \max_{\boldsymbol{x}} \left| \boldsymbol{F}_{\boldsymbol{wv}}{}^{\boldsymbol{c}}(\boldsymbol{x}) - \boldsymbol{F}^{\boldsymbol{c}}(\boldsymbol{x}) \right|$. The significance level of an observed value $\boldsymbol{K^*}$ can be obtained using a numerical approximation (Press *et al.*, 1992).

A gene subset determined by a word pair was deemed to have significantly different expression if its K-S p-value was less than a certain threshold. To correct for multiple testing, this threshold was established by controlling the False Discovery Rate. The significance levels $\boldsymbol{p_i}$ from each K-S test were ordered in ascending order. Let $\boldsymbol{N}$ represent the total number of K-S tests performed, i.e. the number of jointly conserved, closely spaced word pairs times the number of gene expression experiments). If $\boldsymbol{k}$ was the largest $\boldsymbol{i}$ such that $\boldsymbol{p_i} < \boldsymbol{i}\alpha \,/\, \boldsymbol{N}$, then the first $\boldsymbol{k}$ word pairs in the ordering were deemed to have a significance level of $p < \alpha$.

We ensured that the K-S p-value for the conserved word pair subset $\boldsymbol{G}_{\mathrm{o}}$ was more significant than subsets $\boldsymbol{G_w}$ or $\boldsymbol{G_v}$ comprised of only one conserved word by computing $\boldsymbol{K}$

for $\boldsymbol{E}_w{}^{\mathrm{c}}$ vs. $\boldsymbol{E}_{\boldsymbol{v}}{}^{c}$, as well as for $\boldsymbol{E}_w{}^{\mathrm{c}}$ vs. $\boldsymbol{E}^{c}$. The marginal improvement of the joint word pair was defined as: $\boldsymbol{K}\,(\boldsymbol{F}_{\mathrm{o}}{}^{\boldsymbol{c}}$ vs. $\boldsymbol{F}^{\boldsymbol{c}}$) – max($\boldsymbol{K}\,(\boldsymbol{F}_{\mathrm{w}}{}^{\boldsymbol{c}}$ vs. $\boldsymbol{F}^{\boldsymbol{c}}$), $\boldsymbol{K}\,(\boldsymbol{F}_{\mathrm{v}}{}^{\boldsymbol{c}}$ vs. $\boldsymbol{F}^{\boldsymbol{c}}$)).

Hierarchical clustering of word pair associations

The $P \times C$ matrix of K-S *p*-values was log-transformed (base 10), and the word pairs contained in *P* were clustered by average-linkage hierarchical clustering using the program Cluster (Eisen *et al.*, 1998). Since the log-transformed K-S *p*-values were all negative, a centered Pearson correlation was used as the similarity metric.

Stepwise linear regression of gene expression

Regression analyses assume that a log-transformed gene expression measurement, $\boldsymbol{E}_{\boldsymbol{gc}}$ for gene $\boldsymbol{g}$ in condition $\boldsymbol{c}$ can be modeled by a linear equation:

$$E_{gc} = \sum_{f} M_{fc} \times S_{gf} + \varepsilon_g$$

where $\boldsymbol{S}_{\boldsymbol{gf}}$ represents the score of a sequence feature $\boldsymbol{f}$ in gene $\boldsymbol{g}$, $\boldsymbol{M}_{\boldsymbol{fc}}$ represents the influence term of the feature $\boldsymbol{f}$ on gene expression in condition $\boldsymbol{c}$, and ε_g is the gene-specific error term. Genome-wide expression data was filtered for a set of 4703 genes with TCRs conserved in three or more *Saccharomyces* genomes. For a certain experimental condition, the list of features was restricted to either two words found in a single word pair template, or to all words found in conserved word pair templates that were significantly associated with gene expression changes in that condition. The score $\boldsymbol{S}_{\boldsymbol{gf}}$ for feature $\boldsymbol{f}$ in a TCR corresponding to gene $\boldsymbol{g}$ was taken as either the number of occurrences in *S. cerevisiae,* or the number of occurrences conserved in three or more *Saccharomyces* genomes. Stepwise linear regression models were fit to genome-wide expression data using the statistical package R. At each iteration, the sequence feature with the largest increase in the *R*-square goodness-of-fit score was added to the model:

$$R^2 = \sum_f \left(M_{fc} \times S_{gf} \right)^2$$

Pairwise interaction terms between sequence features $\boldsymbol{f_1}$ and $\boldsymbol{f_2}$ already selected in the model, expressed as $\boldsymbol{S_{gf1}} \bullet \boldsymbol{S_{gf2}}$, could also be added to the model at each iteration if the features were found in the same conserved word pair template. Sequence features were added to the regression model as long as the *p*-values for their associated influence terms ($\boldsymbol{M_{fc}}$) were less than 0.05.

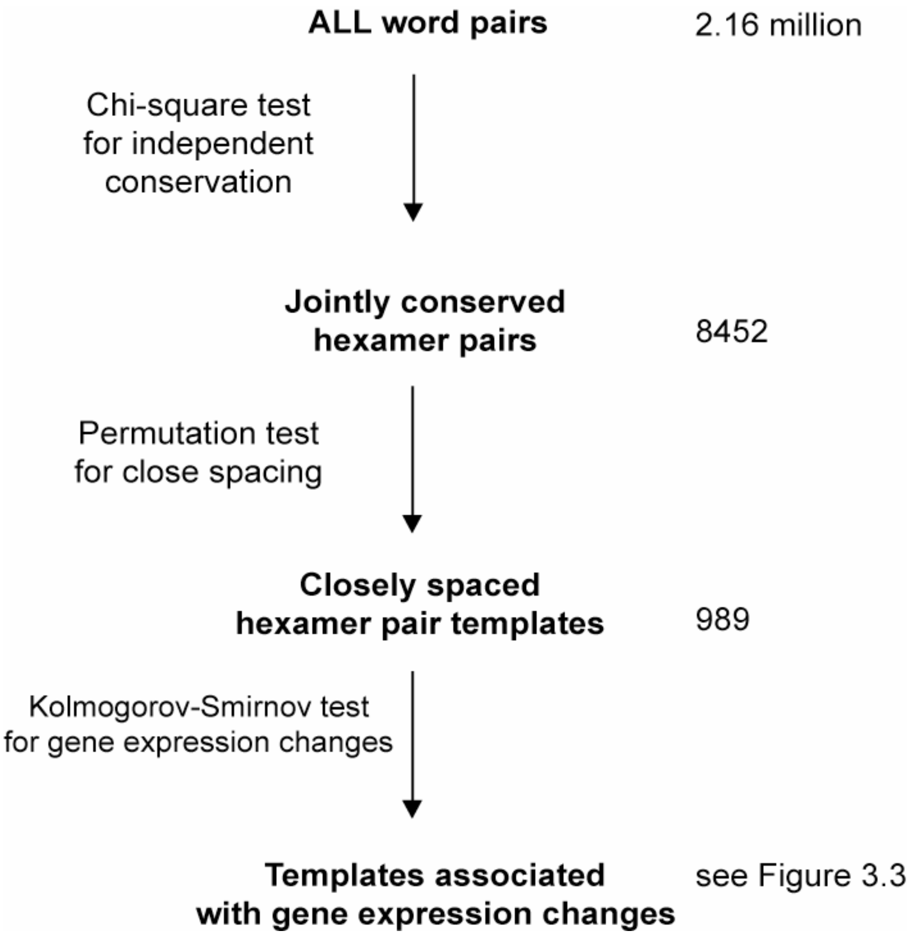
RESULTS

Identification of conserved word pair templates

Multiple genome sequences provide additional power to studies of gene regulation. Due to natural selection, mutations accumulate more rapidly in non-functional DNA regions than in functionally constrained bases. Given a multiple sequence alignment of orthologous sequences from closely related species, the aligned and invariant regions should be enriched for functionally important residues (Duret and Bucher, 1997; Pennacchio and Rubin, 2001). Additional *Saccharomyces* genomes were sequenced to ensure sufficient sequence similarity to *S. cerevisiae* such that orthologous regions could be reliably aligned, yet enough sequence divergence that functional sequences would be much more conserved than non-functional sequences (Cliften *et al.*, 2001; Kellis *et al.*, 2003). In order to confirm that regulatory sequences were found in conserved regions, we tested a database of 47 known, non-redundant regulatory motifs and found that 35 of them show conservation ratios that were more than 3 standard deviations above that expected by random chance (Kellis *et al.*, 2003; Zhu and Zhang, 1999).

We present a method to find conserved higher-order sequence templates from related *Saccharomyces* genomes (Figure 3.1). Our method incorporates sequential statistical tests, with each step focusing on a distinct property of conserved sequence templates. The simplest instances of sequence templates involve word pairs and their relative spacing. As described in detail below, pairs of words that were conserved in the same intergenic regions of four *Saccharomyces* genomes were identified using a chi-square test for independence. Next, a permutation test was used to select word pairs

Figure 3.1) Overview of method to discover conserved word pair templates in yeast



whose physical proximity was closer than that expected by chance. Finally, to evaluate the transcriptional information contained in conserved word pairs with close spacing, the expression of genes containing TCR templates were compared to the rest of the genome. We initialized our word list using all 2080 words of length six, treating a given word and its reverse complement as identical. For each TCR (consisting up to 600 bp upstream of an open reading frame), a word was labeled conserved if all six bases were identical in at least three of the four *Saccharomyces* genomes, based on the CLUSTALW alignment of that TCR.

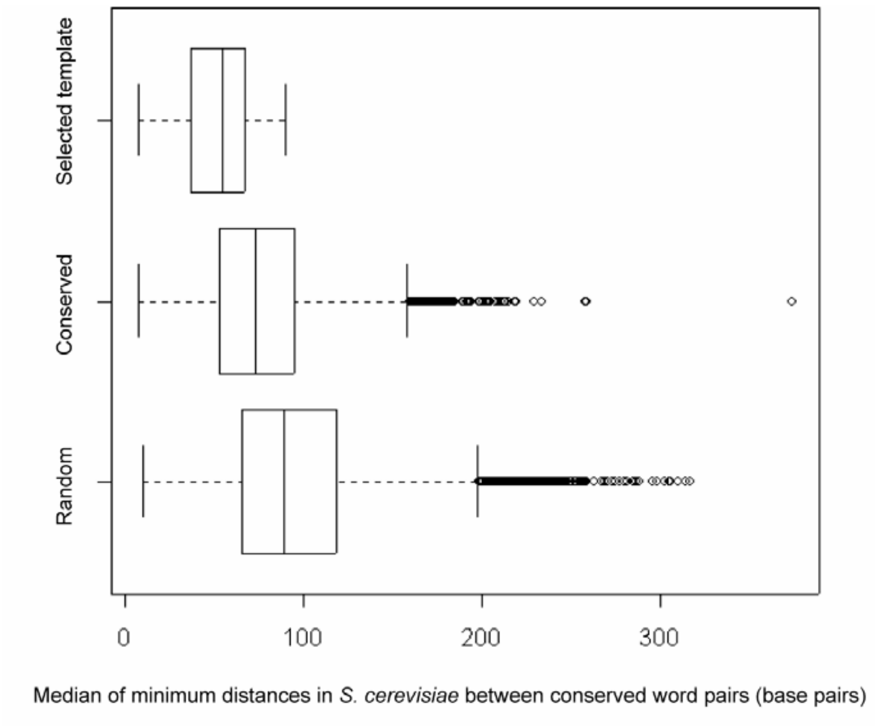
To systematically test whether words were conserved more often in the same intergenic regions of the *Saccharomyces* genomes than expected by independent conservation, a chi-square test was performed on all possible pairwise combinations of words (see Materials and Methods). Pairs of words that overlapped each other by more than three nucleotides were excluded. A significant proportion of word pairs showed dependent conservation: among the 2.16 million word pairs tested, 8452 of them (~0.4%) had conservation χ^2 scores greater than 31.1. This threshold corresponds to a probability of 0.05 for obtaining one or more false positives after a Bonferroni correction for multiple testing.

Next, we selected word pairs that displayed closer physical spacing in intergenic regions than expected by chance. The choice of a statistical test to evaluate close distances must consider the local fluctuations of A+T nucleotide content in genome sequences. Previous work used a Poisson distribution to evaluate proximity between binding sites (Wagner, 1999). However, variability in base composition can skew occurrences of arbitrary sequences away from their expected distributions. Indeed, this statistical test was confounded by large fluctuations in the Poisson parameter estimates, which varied up to 2-fold within a single chromosome (Wagner, 1999).

The effects of base composition fluctuations, as well as varying lengths of TCRs, motivated our nonparametric statistical test for close spacing. We used the median, denoted by $\overline{D}$, to summarize a distribution of minimum distances between two words in *S. cerevisiae*. This distribution was calculated based on the genes whose TCR's conserved both words, and is independent of the relative word ordering. If two non-overlapping words were closely spaced in all TCR's, we should find $\overline{D}$ to be smaller than expected by chance. The statistical significance of this spacing was assessed using a permutation test by selecting the set of genes that contained a conserved word pair and then randomizing the assignment of one of the words to the genes containing that word (see Materials and Methods). By permuting the TCR labels for one of the words, but not the word positions themselves, we retained the positional biases of individual words within intergenic regions.

After correcting for multiple testing by controlling the False Discovery Rate (FDR), a total of 989 out of 8452 word pairs (~12%) had significantly small values (FDR $q < 0.05$) for $\overline{D}$ (Figure 3.2). As a negative control, we also assayed a sample of word pairs that did not show dependent conservation (conservation $\chi^2 < 1$), yet were jointly conserved in at least 10 TCRs. No word pairs in a random sample of 42718 pairs with non-dependent conservation ($\chi^2 < 1$) showed significantly small values for $\overline{D}$, after correction for multiple testing. Figure 3.2 illustrates the distributions of $\overline{D}$ for conserved word pair templates, jointly conserved word pairs, and randomly conserved word pairs.

Figure 3.2) Word pairs in conserved word pair templates are closely spaced in *S. cerevisiae*



A comparison of the median of minimum distances $\overline{D}$ is shown for three categories of word pairs. For each category, the distribution of median of minimum distancess is represented by a box-and-whisker plot, which was generated using the statistical software package R; the box extends from the 25th percentile to the 75th percentile, and the vertical line within the box denotes the median of the distribution. Dashed lines extend for 1.5 times the range of the box, and circles indicate extreme values. “Selected template” denotes closely spaced and jointly conserved word pairs ($\chi^2 > 31.1$, spacing $q < 0.05$, $N = 989$). “Conserved” denotes dependently conserved word pairs that occur in at least 10 intergenic regions ($\chi^2 > 31.1$, $N = 3726$) and includes all of the word pairs in the “selected template” category. “Random” denotes a sample of randomly conserved word pairs that occur in at least 10 intergenic regions ($\chi^2 < 1$).

The medians of these distance distributions were 54 nucleotides, 73 nucleotides and 89 nucleotides, respectively. Notably, the median $\overline{D}$ for template pairs was significantly smaller ($p < 0.05$) than the median $\overline{D}$ for randomly conserved pairs. These results indicate that many of the word pairs that were conserved in the same intergenic regions of multiple *Saccharomyces* genomes also exhibited closer spacing in TCRs.

Conserved word pair templates were significantly associated with gene expression

Our method identified conserved word pair templates that were statistically significant with respect to both co-conservation in multiple genomes and close spacing in *S. cerevisiae* TCRs. To evaluate the regulatory information in these templates, we assessed the statistical association between gene groups that shared a template and changes in gene expression. Similar to other group-by-sequence approaches for finding regulatory sequences, we expect that gene subsets defined by common TCR sequence features should have gene expression patterns that are similar under conditions where the transcription factors are active, yet are different from the average expression of genes in the genome (Chiang *et al.*, 2001).

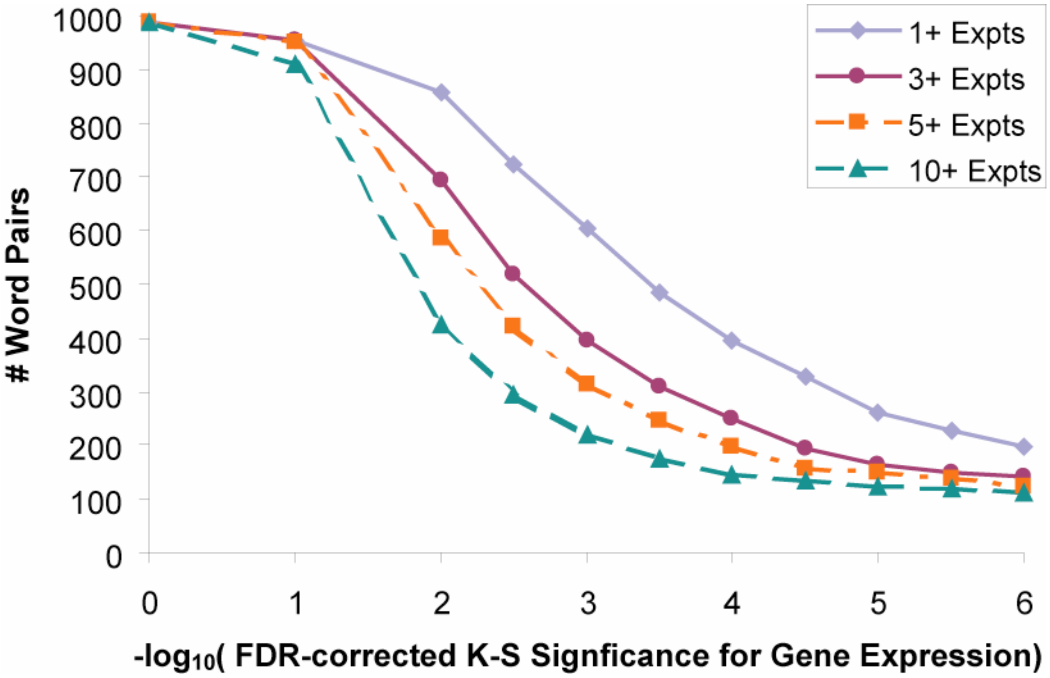
To assess the association between conserved word pair templates and differentially expressed genes, we identified gene subsets that contained both conserved words in the template within their TCRs and observed their expression patterns in *S. cerevisiae* in publicly available datasets (Gasch *et al.*, 2000; Spellman *et al.*, 1998; Cho *et al.*, 1998; Gasch *et al.*, 2001; Lee *et al.*, 2000; Hughes *et al.*, 2000b; see Materials and Methods for details). We then conducted Kolmogorov-Smirnov (K-S) tests to evaluate for differential gene expression between each gene subset and the whole genome. K-S tests provide a nonparametric, sensitive and robust way to compare two distributions.

Similar results were obtained using other statistical tests, such as *t*-tests and likelihood ratio tests (A. M. Moses, unpublished data). A $P \times C$ matrix was computed: each conserved word pair in *P* was assigned a K-S *p*-value for each experimental condition observed in *C*. (see Materials and Methods). Entries in this matrix (K-S *p*-values) were filtered out if the K-S *p*-value: (1) did not meet the threshold for multiple testing; or (2) was less than 10 times more significant than the K-S *p*-value for a gene subset associated with either word alone (see Materials and Methods). The latter criterion minimizes gene expression changes that can be explained by the presence of a single conserved word.

Figure 3.3 displays the number of conserved word pair templates that were significantly associated with gene expression changes, for varying significance levels of the K-S test, which have been corrected for multiple testing (see Materials and Methods). Each line indicates the number of gene subsets that were significant in a different minimum number of experimental conditions. Several hundred closely spaced word pairs were significantly associated with differential gene expression. For example, 314 word pairs met an FDR-corrected significance threshold of $p < 10^{-3}$ for 5 or more experimental conditions, which represented 32% of all closely spaced word pairs.

The proportion of conserved word pair templates showing significant associations with gene expression was compared to two sets of negative controls, comprising word pairs that failed either the first (co-conservation) or second (close spacing) statistical test. As the first control, we used a sample of 624 word pairs that failed the joint conservation test (conservation $\chi^2 < 1$) found in at least 25 TCRs, but also showed modest constraints on word pair spacing ($p < 0.15$). Only 8 of these word pairs (1.3%) had significant

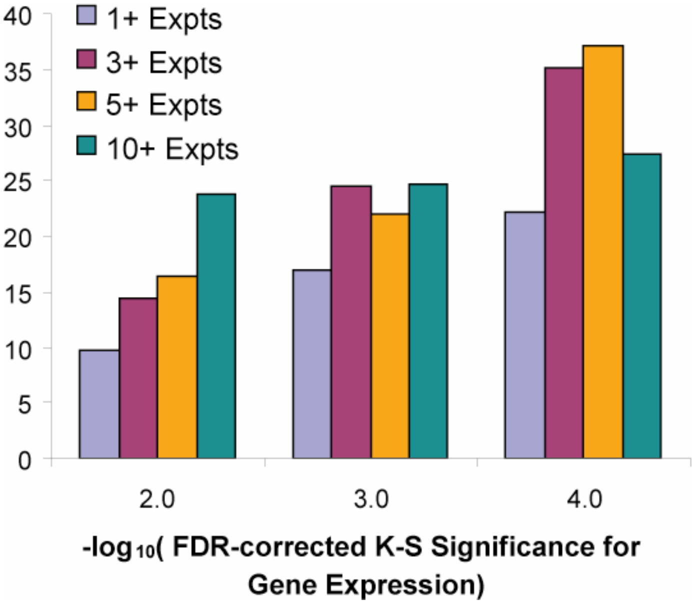
Figure 3.3) Total number of conserved word pair template associations at different K-S significance values



The horizontal axis shows different False Discovery Rate-corrected significance levels for the Kolmogorov-Smirnov test (see Materials and Methods). The number of closely spaced word pairs meeting this cutoff for different minimum numbers of expression conditions is shown on the vertical axis. Word pairs were also filtered for an improvement of 10× over the K-S significance from any single word.

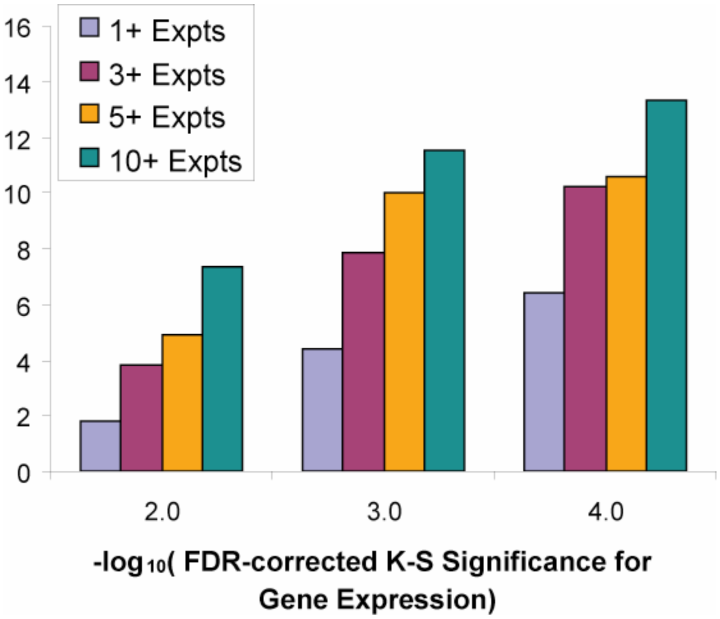
expression at an FDR-corrected threshold of $p < 10^{-3}$ for 5 or more experimental conditions. To assess the relative enrichment for significant associations with gene expression changes at a variety of multiple testing thresholds, we computed an odds ratio: the proportion of significant associations among the template pairs, divided by the proportion of significant associations among the random pairs. For the above threshold, the odds ratio was about 22. In other words, gene groups that contain a common conserved word pair template in their TCRs were about 22 times more likely to be associated with significant gene expression changes, compared with gene groups selected using randomly conserved word pairs. As shown in Figure 3.4, the odds ratios for association with gene expression changes in multiple conditions varied between 10 and 35. This analysis was repeated for a sample of 2737 co-conserved (conservation $\chi^2 > 31.1$) word pairs that failed the close spacing test (permutation $p > 0.05$ after multiple testing), yet occurred in at least 10 intergenic regions. The relative enrichment for gene expression associations in closely spaced words is displayed in Figure 3.5. Among co-conserved word pairs, those pairs that were closely spaced than expected were still about 2 to 12 times more likely to be significantly associated with gene expression changes, compared to word pairs that were not found to have significantly close spacing. We confirmed that gene groups associated with significant gene expression changes did not have statistically significant differences in their TCR sizes, as assessed by a permutation test (data not shown). Thus, gene groups that contained co-conserved and spatially close word pairs are more significantly associated with gene expression changes than expected by chance.

Figure 3.4) Relative enrichment for significant gene expression associations compared to independently conserved words.



Relative enrichment was computed as an odds ratio: the fraction of gene groups selected by conserved word pair templates associated with significant gene expression changes, divided by the fraction of gene groups selected by randomly conserved word pairs associated with significant gene expression changes. Templates were chosen as the set of 989 word pairs showing dependent conservation and close spacing ($\chi^2 > 31.1$, spacing $q < 0.05$); the random word pairs included 624 pairs showing independent conservation and modest spacing constraints ($\chi^2 < 31.1$, spacing $q < 0.15$). The odds ratio is shown on the vertical axis; various FDR-corrected significance thresholds for gene expression associations are shown on the horizontal axis. Word pairs were filtered for an improvement of 10× over the K-S significance from any single word.

Figure 3.5) Relative enrichment for significant gene expression associations compared to co-conserved words that failed the close spacing test



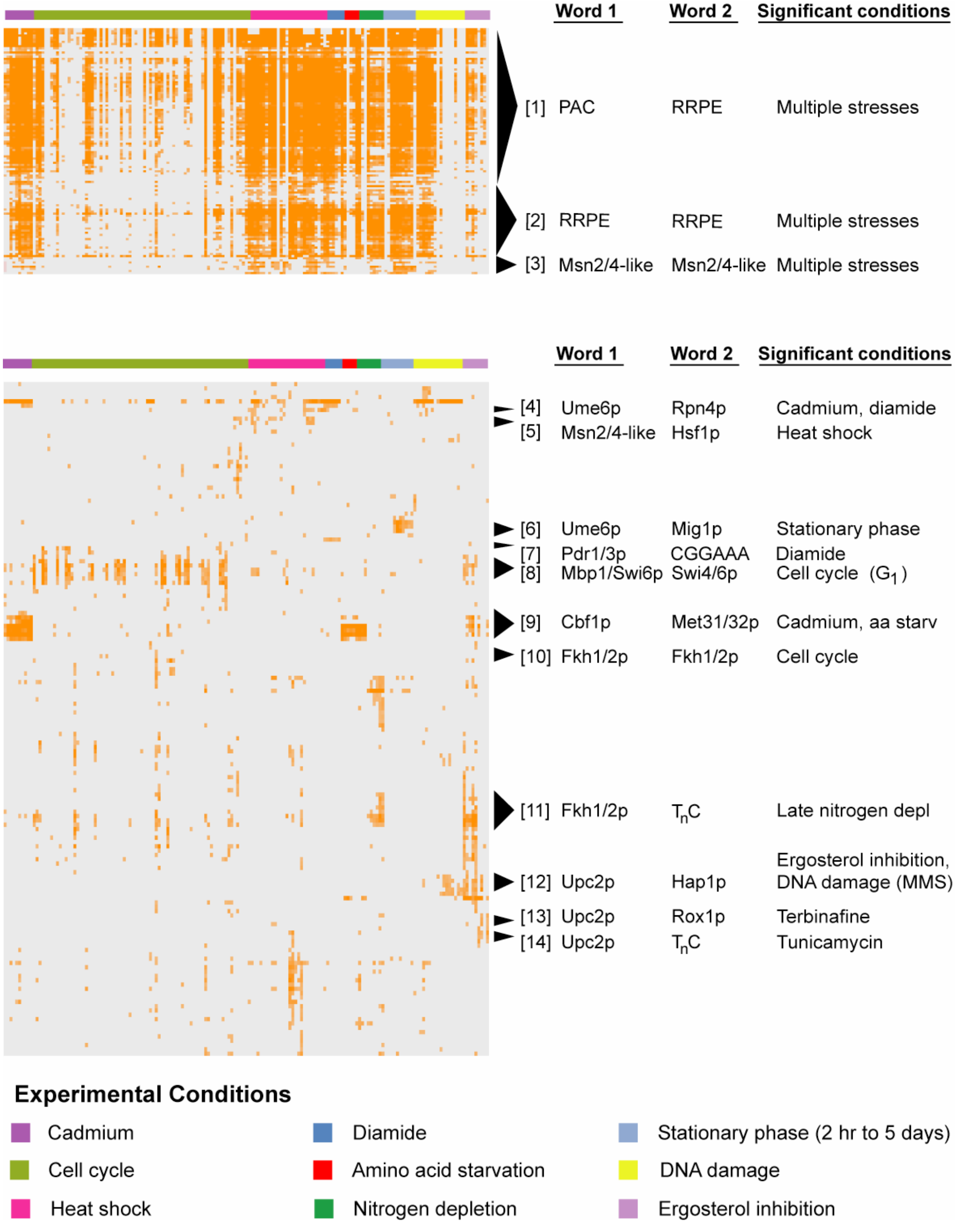
Templates were chosen as the set of 989 word pairs showing dependent conservation and close spacing ($\chi^2 > 31.1$, spacing $q < 0.05$); the background word pairs included 2737 pairs showing co-conservation, but no significant close spacing constraints ($\chi^2 > 31.1$, spacing $q > 0.05$). The odds ratio is shown on the vertical axis; various FDR-corrected significance thresholds for gene expression associations are shown on the horizontal axis. Word pairs were filtered for an improvement of 10× over the K-S significance from any single word.

Many identified sequences represented known transcription factor binding sites

In addition to their statistical significance, many conserved word pair templates that were most strongly associated with gene expression changes were consistent with biological information on the transcription factors known to bind those sites (Costanzo *et al.*, 2001). In all analyses described below, we used 314 word pairs that had significant associations with gene expression changes at an FDR-corrected multiple testing threshold of $p < 10^{-3}$ for 5 or more experiments. For visualization purposes, we organized the $P \times C$ matrix by hierarchically clustering the K-S *p*-values for the 314 word pairs

Hierarchical clustering of this output matrix identified groups of word pairs with similar K-S *p*-values in specific subsets of experimental conditions (Figure 3.6). In many cases, the word pairs that clustered together also comprised overlapping hexamer sequences, suggesting that some of the hexamers in different pairs may represent a larger, somewhat variable sequence (Table 3.1). For example, group #9 in Figure 3.6 included 6 word pairs. In each of these word pairs, one of the component words – such as TCACGT, CACGTG, or ACGTGC – matched part of the Cbf1 consensus binding site (TCACGTG). The other component word in each pair – such as ACTGTG, CTGTGG, TGTGGC or GTGGCT – represented part of the known Met31 or Met32 binding site (AAACTGTGG). Therefore, genes whose TCRs contained any word pair within this group likely contained a conserved Cbf1 binding site, along with a conserved Met31 or Met32 binding site, and the distances between the conserved sites in these genes were also smaller than expected by chance. These results agree with the known interaction of Cbf1 and Met31 or Met32 for the regulation of genes involved in sulfur utilization (see Discussion).

Figure 3.6) Specific patterns of gene expression changes are associated with templates



	Word 1	Word 2	Significant conditions
[1]	PAC	RRPE	Multiple stresses
[2]	RRPE	RRPE	Multiple stresses
[3]	Msn2/4-like	Msn2/4-like	Multiple stresses

	Word 1	Word 2	Significant conditions
[4]	Ume6p	Rpn4p	Cadmium, diamide
[5]	Msn2/4-like	Hsf1p	Heat shock
[6]	Ume6p	Mig1p	Stationary phase
[7]	Pdr1/3p	CGGAAA	Diamide
[8]	Mbp1/Swi6p	Swi4/6p	Cell cycle (G$_1$)
[9]	Cbf1p	Met31/32p	Cadmium, aa starv
[10]	Fkh1/2p	Fkh1/2p	Cell cycle
[11]	Fkh1/2p	T_nC	Late nitrogen depl
[12]	Upc2p	Hap1p	Ergosterol inhibition, DNA damage (MMS)
[13]	Upc2p	Rox1p	Terbinafine
[14]	Upc2p	T_nC	Tunicamycin

Experimental Conditions

■ Cadmium
■ Cell cycle
■ Heat shock
■ Diamide
■ Amino acid starvation
■ Nitrogen depletion
■ Stationary phase (2 hr to 5 days)
■ DNA damage
■ Ergosterol inhibition

Figure 3.6 (continued)

The $P \times C$ matrix of K-S *p*-values was hierarchically clustered by rows and visualized with TreeView (http://rana.lbl.gov). Each row corresponds to a conserved word pair template, and each column represents a single gene expression experiment. The experimental conditions are indicated by the color bar above and below the figure, according to the key shown below. The value in each cell corresponds to the K-S *p*-value of gene expression changes in each condition (column) for a group of genes that contain the conserved word pair template (row) in their TCRs. An orange color denotes a K-S *p*-value below the FDR critical value of 0.001 for multiple testing, while grey represents values that were not significant. Word pairs that failed to meet a False Discovery Rate critical value of 0.001 for multiple testing in 5 or more experiments are not shown. Some of the most significant conserved word pair associations are labeled and annotated in Tables 3.1 and 3.2.

Table 3.1) Consensus sequences for the most significant groups of word pairs

	Hexamer list for word 1	Compiled sequence 1	TF for consensus 1	Hexamer list for word 2	Compiled sequence 2	TF for consensus 2	# word pairs
1	**G**A**GATG** **G**C**GATG** A**GATGA** C**GATGA** **GATGAG** **ATGAG**A **ATGAG**C **TGAG**A**T** **TGAG**C**T** **GAG**A**T**G **AG**A**T**GA **AG**C**T**CA	**GMGATGAGMTSA**	PAC motif (Hughes *et al.*, 2000a)	**TGAAAA** **GAAAAA** **AAAAAT** **AAAATT** **AAATTT**	**TGAAAATTT**	RRPE motif (Hughes *et al.*, 2000a)	75
2	**AA**G**TGA** **A**A**TGAA** **A**G**TGAA** A**TGAAA** C**TGAAA** **TGAAAA**	**A**N**TGAAAA**A	RRPE motif (Hughes *et al.*, 2000a)	**GAAAA**A **GAAAA**T **AAAA**T**T** **AAA**T**T**T	**GAAAAWT**T	RRPE motif (Hughes *et al.*, 2000a)	40
3	GTT**CCC** CT**CCCC** A**CCCCT** T**CCCCT**	GYW**CCCCT**	Msn2/4-like (Discovered Motif 38 Kellis *et al.*, 2003)	CCCTTT CCTTTT CCTTAT	C**CCTTWT**	Msn2/4-like (Discovered Motif 38 Kellis *et al.*, 2003)	5
4*	G**GCGGC** **GCGGC**T	G**GCGGC**T	Ume6	GT**GGCA** **GGCA**AA	GT**GGCA**AA	Rpn4	2
5	C**CCTTT** **CCTTT**T	C**CCTTT**T	Msn2/4-like	G**G**A**GAA** **G**G**GAA**A	G**G**R**GAA**A	Hsf1	2

	Hexamer list for word 1	Compiled sequence 1	TF for consensus 1	Hexamer list for word 2	Compiled sequence 2	TF for consensus 2	# word pairs
6	**CGGCGG**	CGGCGG	Ume6	T**ACCCC** **ACCCCA** **CCCCA**A	T**ACCCCA**A	Mig1	3
7*	CCGCGG	CCGCGG	Pdr1/3	CGGAAA	CGGAAA	Unknown	1
8	AA**ACGC** G**ACGCG** A**ACGCG** **ACGCG**T **ACGCG**A T**CGCG**T **CGCG**TC	A**RWCGCG**W	Swi6/ Mbp1	CG**CGAA** A**CGAAA** G**CGAAA** **CGAAA**C **CGAAA**A	C**RCGAAA**M	Swi4/6	9
9	T**CACGT** **CACGT**G **ACGTG**C	T**CACGTG**C	Cbf1	A**CTGTG** **CTGTGG** **TGTGGC** **GTGGC**T	A**CTGTGGC**T	Met31 or Met32	6
10	TA**T**T**TT** **T**T**TTGT** T**TTGTT** A**TTGTT**	**TWTTGTT**	Fkh1/2	T**GTTTA** **GTTTA**C	T**GTTTA**C	Fkh1/2	4
11	T**TTGTT** **TTGTTT**T	T**TTGTTT**T	Fkh1/2	**TTTTT**C **TTTTT**T	**TTTTT**Y	T_nC	4
12 *	T**CGTTT** **CGTTT**A	T**CGTTT**A	Ecm22 | Upc2	C**CGATA** **CGATA**A	C**CGATA**A	Hap1	4
13	T**CGTTT** **CGTTT**A	T**CGTTT**A	Ecm22 | Upc2	T**ATTGT** **ATTGT**T	T**ATTGT**T	Rox1	2
14	C**GTTTC** **GTTTC**T	**CGTTTC**T	Ecm22 | Upc2	T**TCTTT** **TCTTTT** **CTTTT**T	T**TCTTTT**T	T_nC	5

Table 3.1 (continued)

The output $\boldsymbol{P} \times \boldsymbol{C}$ matrix of word pairs ($\boldsymbol{P}$) that were significantly associated ($p < 0.001$) with at least 5 or more environmental conditions ($\boldsymbol{C}$) was ordered using hierarchical clustering. Numbers correspond to groups of overlapping word pairs indicated in Figure 3.6. Stars denote sequence pairs whose involvement in multifactorial regulation has not been previously reported. Compiled sequences were assembled from groups of word pairs that were found in adjacent rows in the ordering of K-S *p*-values. Since individual words must have passed all three statistical tests to be included in the output matrix, these consensus sequences may not reflect the actual biological specificities of conserved transcription factor binding sites (refer to Kellis *et al.*, 2003 for a more complete list). Residues are shown in bold if it is contained in at least two hexamers. Numbers denote the groups that are indicated in Figure 4. IUPAC codes used: K (G or T); M (A or C); R (A or G); S (C or G); W (A or T).

Table 3.1 shows a partial list of the 14 most significant groups of consensus sequences, which were assembled by joining adjacent word pairs in the clustered output matrix with overlapping sequences. Many of these consensus sequences matched transcription factor binding sites that had been biochemically verified. Several pairs of transcription factors, denoted by asterisks in Table 3.1, were not previously known to act on the same sets of target genes.

Conditions with significant gene expression changes coincided with transcription factor activity

Further support that templates contain transcriptional regulatory information was obtained from a key observation: the experimental conditions with significant gene expression changes often corresponded to conditions in which the cognate transcription factors are known to be active (Table 3.2). In addition, many gene subsets that shared an individual word pair template in their TCRs were highly enriched for gene expression changes. We will survey examples of word pair templates associated with gene expression changes, focusing our attention on several environmental stress conditions. The environmental stress response represents a paradigm for multifactorial control of transcription regulation. Genome-wide expression studies found that ~300 genes were induced and ~600 genes were repressed in response to a wide variety of stressful environmental transitions (Gasch *et al.*, 2000; Causton *et al.*, 2001). Many of these genes also showed subtly different expression patterns in response to specific stimuli, suggesting that the common environmental stress response may be modulated by the activity of condition-specific transcription factors (Gasch *et al.*, 2000).

Table 3.2) Summary for most significant groups of conserved word pairs

	Conserved Word Pairs (Compilation of overlapping words)	Known transcription factors or motifs	Conservation (χ^2, *p*-val via Bonferroni)	Median of min dist $\overline{D}$	# TCR	Expression conditions with significant gene subsets (FDR significance)
1	G[AC]GATGAG TGAAAATTTT	PAC, RRPE	240.6 (10^{-49})	19 ± 0.5	162	Repressed in multiple environmental stresses (10^{-6})
2	ANTGAAA, GAAAAWT	RRPE (Overlap)	96.9 (2×10^{-16})	43 ± 11	68	Repressed in multiple environmental stresses (10^{-6})
3	**CTCCCC,** **CCCTTA**	Msn2/4-like, (Overlap)	53.8 (5×10^{-7})	28 ± 3.7	15	Induced in multiple environmental stresses (10^{-6})
4	GGCGGC, GTGGCA	Ume6, Rpn4	43.7 (9×10^{-5})	48 ± 16	25	Cadmium, diamide (10^{-4}) MMS, heat shock (10^{-3})
5	**CCTTTT,** GAGAAA	Msn2/4, Hsf1	56.2 (2×10^{-7})	54 ± 5.4	69	Heat shock (10^{-4})
6	CCGCCG, ACCCCA	Ume6, Mig1	41.9 (2×10^{-4})	17 ± 1.5	14	Stationary phase (10^{-6})
7	**CCGCGG,** CGGAAA	Pdr1/3, Unknown	111 (2×10^{-19})	44 ± 12	21	Diamide (10^{-3})
8	**RACGCG,** **RCGAAA**	Swi6/Mbp1, Swi4/6,	83.0 (7×10^{-13})	33 ± 5.0	33	Cell cycle, G1 phase (10^{-6})
9	GCACGTGC, ACTGTGGC	Cbf1 | Pho4, Met31 or Met32	37.4 (2×10^{-3})	22 ± 2.5	22	Amino acid starv. (10^{-6}) Nitrogen depletion (10^{-6}) Cadmium (10^{-6})
10	**T[AT]TTGTT** TGTTTAC	Fkh1/2 (Overlap)	51.1 (2×10^{-6})	57 ± 6.9	48	Cell cycle (10^{-3})
11	TTTGTT, TTTTTY	Fkh1/2, T_nC	37.6 (2×10^{-3})	49 ± 4.4	267	Late nitrogen depletion (10^{-3})

12	CCGATA, TCGTTT	Hap1, Ecm22 | Upc2	36.2 (4×10^{-3})	41 ± 5.9	28	Ergosterol inhibition (10^{-4}) MMS (DNA damage) (10^{-3})
13	TCGTTT, TATTGTT	Rox1, Ecm22 | Upc2	58.8 (4×10^{-8})	55 ± 0.5	69	Ergosterol inhibition (10^{-3}) Early menadione (10^{-3})
14	TGACTC, TCTTTT	Gcn4, T_nC	35.6	59 ± 9.1	63	Ergosterol inhibition (10^{-5}) Amino acid starvation (10^{-5})

Statistics are listed for one representative word pair for each group of overlapping word pairs, numbered as in Figure 3.6. Multiple transcription factors that may bind the same sequence motif are separated by "|". To summarize the close spacing ($\overline{D}$) between conserved word pairs, we report the median of the distribution of minimum distances in *S. cerevisiae* ± standard deviation of the medians of the distribution of minimum distances in all four *Saccharomyces* genomes.

Over one-third of the conserved word pair templates were associated with gene expression changes in multiple environmental stress conditions (Figure 3.6, Tables 3.1 and 3.2). The largest group of overlapping word pairs contained matches to the PAC and RRPE motifs, which were associated with genes that were repressed in multiple stresses (Gasch *et al.*, 2000; Hughes *et al.*, 2000a). These motifs were discovered by their enrichment among the approximately 600 genes that were commonly repressed in stress, yet the putative transcription factors that bind these sequences have yet to be determined. The second largest group of overlapping word pairs corresponded to the RRPE core, which is 10 nucleotides long, along with some flanking conserved bases. These repressed genes were enriched for rRNA processing genes, the group of genes in which this motif was originally identified (Hughes *et al.*, 2000a). Nine conserved word pair templates contained sequences that matched most of the stress response element (STRE), the consensus site for the general stress transcription factors Msn2/Msn4. Genes that conserved both of these words in their TCRs were significantly associated with gene expression induction in multiple environmental stresses, including cadmium, heat shock, amino acid starvation, nitrogen depletion, and stationary phase. In most cases, the sequences comprising the word pairs were mutually overlapping. We interpret these sequences as representing different halves of the same binding site. Since our test for close spacing required non-overlapping sequences, the two words must have appeared over 6 bp away in TCRs. Thus, these genes have likely conserved at least two Msn2/4-like consensus sequences in their TCRs.

Several groups of conserved word pair templates only showed significant associations with gene expression in different subsets of stress conditions (Figure 3.6, Tables 3.1 and 3.2). For example, binding sites for Cbf1 and Met31 or Met32 were found to co-occur in several conserved word pair templates. Genes that contained conserved binding sites for these transcription factors in their TCRs were strongly induced in cadmium, amino acid starvation and early nitrogen depletion. These conditions are consistent with the biological activity of these transcription factors, which induce transcription of sulfur assimilation genes in response to the demand of sulfur-containing metabolites (Thomas and Surdin-Kerjan, 1997; Blaiseau and Thomas, 1998; Fauchon *et al.*, 2002). In another example, several word pairs comprising binding sites for the transcriptional repressors Mig1 and Ume6 were associated with induced gene expression in stationary phase. The observed derepression of Mig1 and Ume6 targets in stationary phase is consistent with the nuclear export of the Mig1 repressor under glucose limitation, as well as recent findings that carbon source genes can be Ume6 targets (Williams *et al.*, 2002). In addition, genes containing a conserved sequence similar to the consensus for Msn2/4, an inducer of the environmental stress response, and the heat-shock transcription factor Hsf1 were significantly induced under heat shock. Once again, the conditions with most significant gene expression changes corresponded to the known activities of the transcription factors involved.

Enrichment for known transcription factor targets among individual gene groups

Some groups of genes with shared word pair templates were enriched for known targets of transcription factors. The vast majority of genes with conserved sites for both the Cbf1 and Met31 or Met32 transcription factors were induced more than 4-fold in cadmium, amino acid starvation, and early nitrogen depletion (Figure 3.7A). Half of

Figure 3.7) Enrichment for known transcription factor targets among individual gene groups

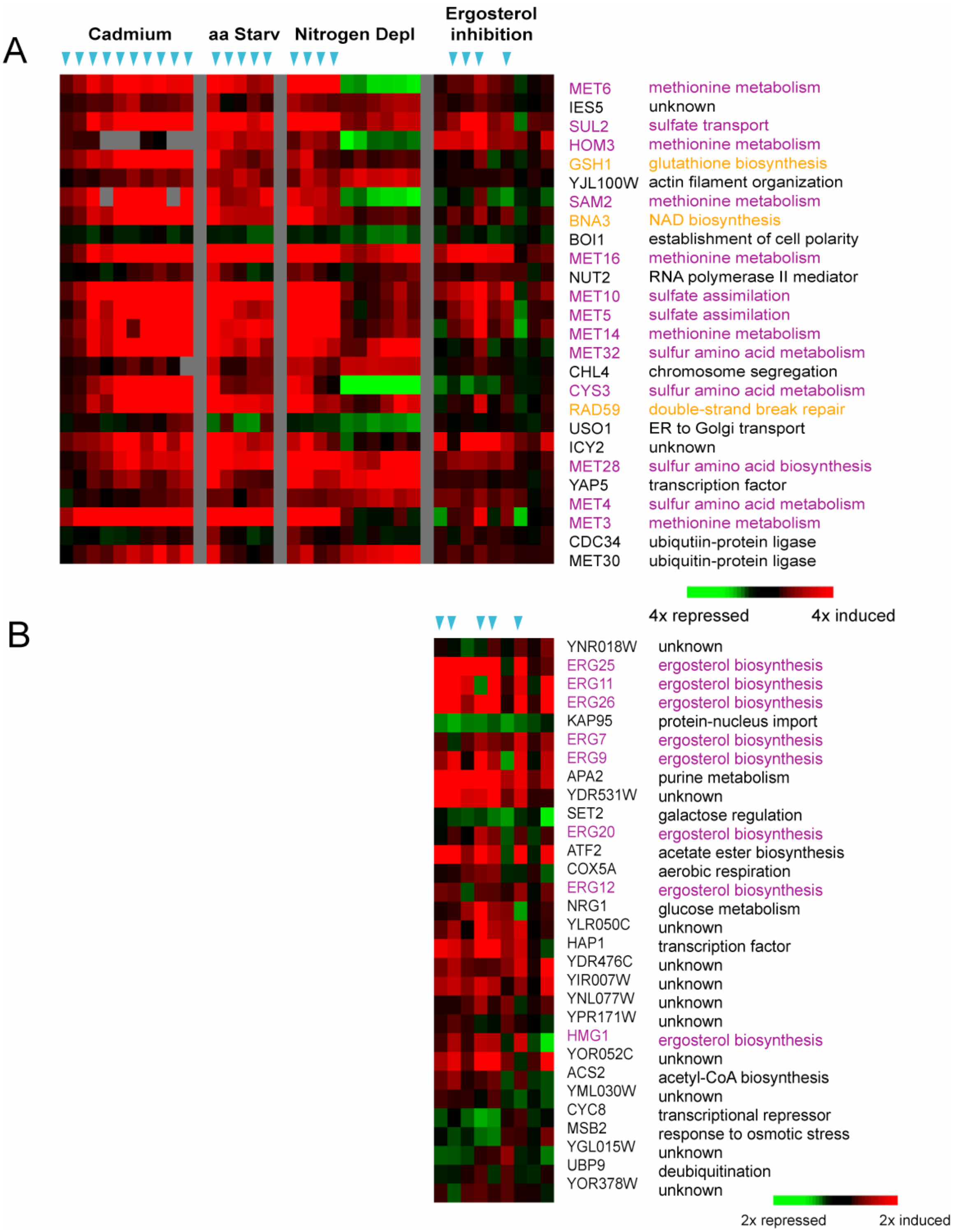


Figure 3.7 (continued)

Gene expression patterns are shown for genes whose TCRs contain the known binding sites for: (A) Cbf1 (CACGTG) and Met31 or Met32 (TGTGGC); or (B) Hap1 (CCGATA) and Ecm22/Upc2 (TCGTTT). The genes are listed in ascending order of minimum distance between the two conserved words in the corresponding TCRs of *S. cerevisiae*. Each row in these diagrams represents a given gene's expression pattern under the conditions shown in each column: exposure to increasing concentrations of cadmium chloride (from 0.05 mM to 0.4 mM); an amino acid starvation timecourse; a nitrogen-source depletion timecourse Gasch *et al.*, 2000; and growth in the presence of drugs or genetic alterations that inhibit ergosterol biosynthesis (*erg3*Δ, itraconazole, *erg28*Δ, overexpressed *ERG11*, *erg2*Δ, tunicamycin, terbinafine, erg6Δ, overexpressed *HMG2*) (Hughes *et al.*, 2000b). A red color indicates that the gene's expression was induced under those conditions, while a green color indicates that the gene was repressed under those conditions; black indicates no detectible change in expression, and grey indicates missing data. Gene names in purple correspond to genes with confirmed roles in (A) sulfur utilization or (B) ergosterol biosynthesis; gene names in orange show highly correlated expression patterns, despite their lack of annotation as sulfur utilization genes. Arrows above the columns indicate conditions in which the displayed gene groups show significant gene expression changes according to the Kolmogorov-Smirnov test, after False Discovery Rate correction for multiple testing at a *p*-value of 0.001.

these genes have confirmed roles in sulfur utilization processes, such as methionine metabolism, sulfate assimilation, sulfate transport and sulfur amino acid metabolism. Compared to the rest of the genome, the group of genes that conserved both of these words within their TCRs was highly enriched for sulfur utilization genes (hypergeometric p-val < 1×10^{-16}, after Bonferroni correction for multiple testing). In addition, we found 3 genes in this group (*GSH1*, *RAD59* and *BNA3*) with highly correlated expression under the above conditions, and thus may be commonly regulated by Cbf1 and Met31 or Met32, despite their lack of direct annotation as sulfur utilization genes. The shared conservation of both the Cbf1 and Met31 or Met32 sites provides further evidence that these genes comprise part of the cellular response to the demand for products of this pathway.

Genes with a conserved half-site for the Hap1 transcription factor, as well as a conserved Ecm22/Upc2 binding site in their TCRs, were significantly associated with induction in the presence of drugs that inhibited ergosterol biosynthesis (Figure 3.7B). This group of 30 genes contained 8 ergosterol biosynthesis genes; this proportion represented an enrichment compared to the rest of the genome (hypergeometric $p < 6 \times 10^{-6}$ after Bonferroni correction for multiple testing). The transcription factors Ecm22 and Upc2 have been shown to induce the expression of ergosterol biosynthesis genes in response to low intracellular concentrations of ergosterol, while Hap1 is known to regulate the expression of these genes according to the availability of heme and oxygen which are required for the pathway (Vik and Rine, 2001; Kwast *et al.*, 1998).

Conserved word pairs were more informative than sequence features derived from single words or single species

The above results from the K-S test strongly suggested that sequence features based on the co-conservation and close spacing of word pairs identified examples of multifactorial regulation. Two other statistical tests were also performed to examine how information contained in conserved word pair templates compared to other sequence features derived from *S. cerevisiae,* or from single conserved words. Informative sequence features should be both highly specific (a high proportion of genes that share the feature should show gene expression changes) and highly sensitive (most of the genes that show gene expression changes should also share the feature).

To assess the specificity of a sequence feature in explaining gene expression, we computed the average expression profile for all genes that shared that feature. We expect that if a sequence feature represented a transcription factor binding site, genes containing that feature in their TCRs would be induced (or repressed), on average, compared to all the genes in the genome (Chiang *et al.*, 2001). By comparing the average expression profile derived from conserved word pair templates with average expression profiles derived from simpler sequence features, we assessed how much information was obtained by incorporating both the conservation and pairwise combination of sequences. For representative word pairs associated with significant gene expression changes in environmental stress conditions, we compared the average expression profile for: gene subsets that shared single words in *S. cerevisiae*; gene subsets that conserved single words among multiple genomes; and gene subsets that shared both words in *S. cerevisiae* (Figure 3.8). In general, the average gene expression profiles for conserved word pairs

Figure 3.8) Incorporation of conservation and word pairs provided more informative average expression profiles

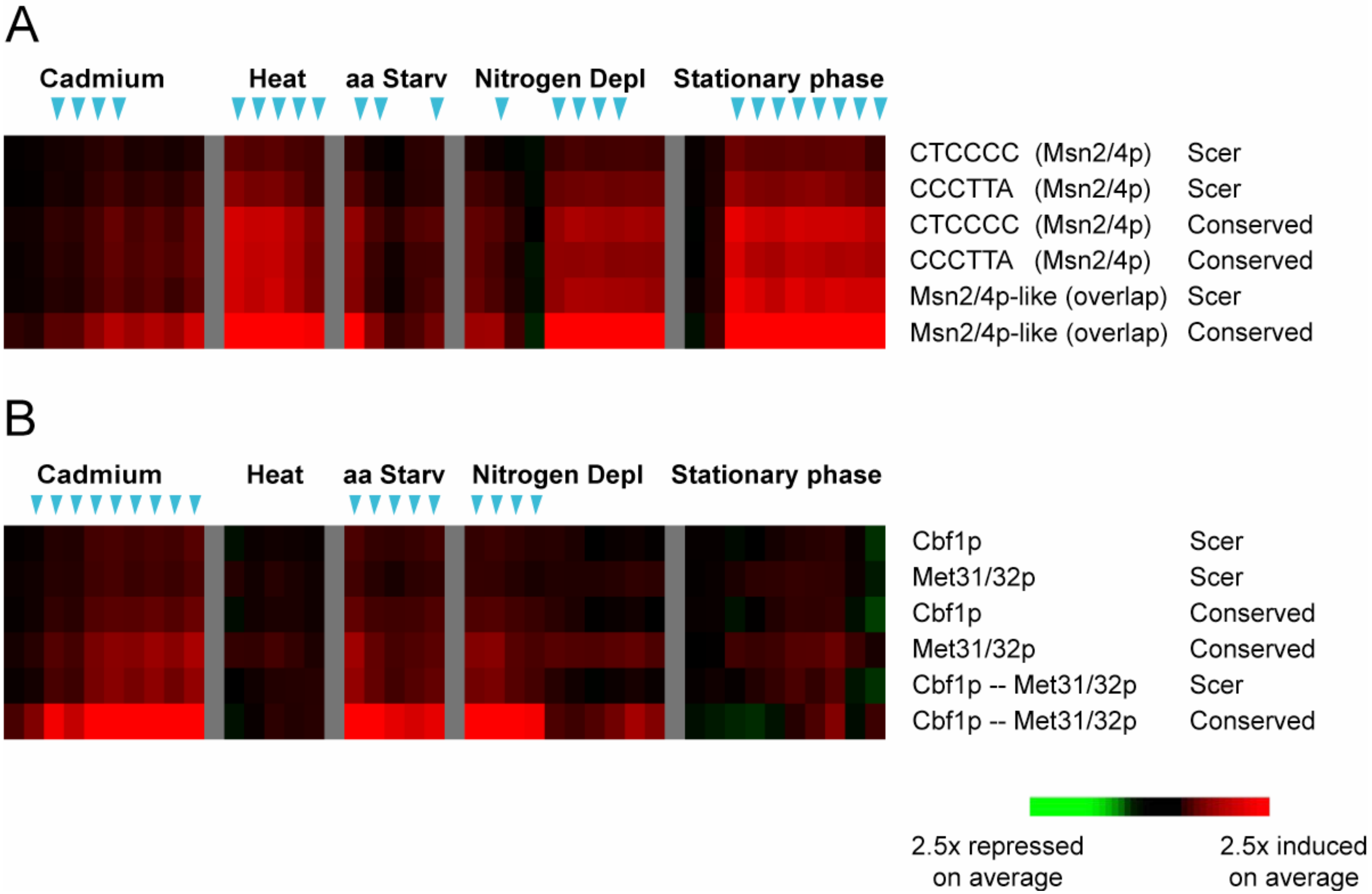


Figure 3.8 (continued)

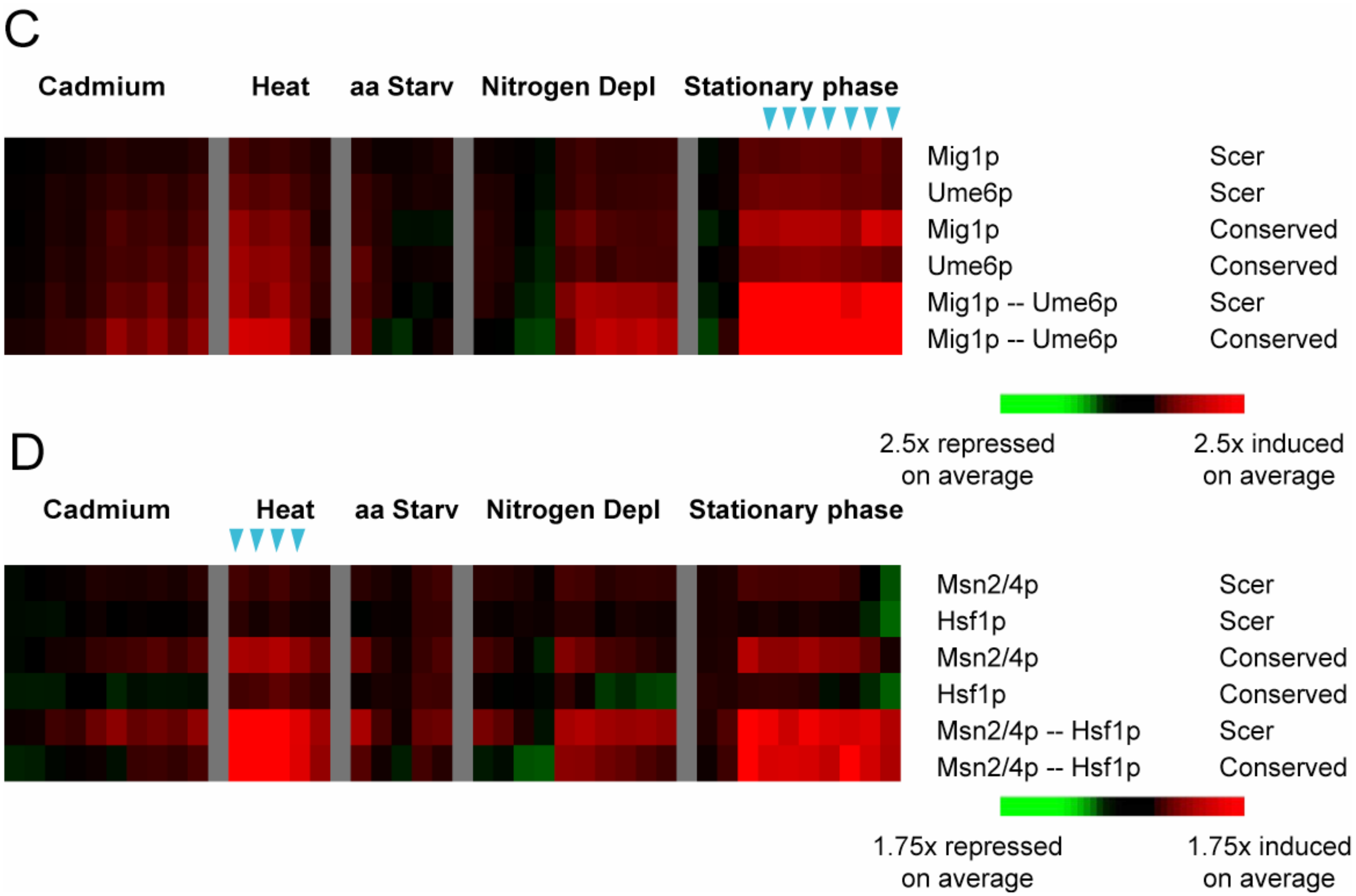


Figure 3.8 (continued)

Groups of genes whose TCRs contained various sequence features were summarized by the average of their gene expression profiles. Each row in these diagrams represents a given gene group's average expression pattern under the conditions shown in each column: exposure to increasing concentrations of cadmium chloride (from 0.05 mM to 0.4 mM); 20 minutes after heat shock to 37°C (from 17°C, 21°C, 25°C, 29°C, and 33°C); an amino acid starvation timecourse; a nitrogen-source depletion timecourse; progression into stationary phase (2 h, 4 h, 6 h, 8 h, 10 h, 12 h, 1 day, 2 days, 3 days, 5 days of growth) Gasch *et al.*, 2000. Representative conserved word pair templates were chosen for analysis, corresponding to: (A) Msn2/4-like sequences (CTCCCC and CCCTTA); (B) Cbf1 (CACGTG) and Met31 or Met32 (TGTGGC) binding sites; (C) Mig1 (ACCCCA) and Ume6 (CCGCCG) binding sites; (D) Msn2/4-like (CCCCTT) and Hsf1-like (GAGAAA) sequences. For each of the panels (A) through (D), each row represents the average expression profile for gene groups chosen by different sequence features in their TCRs: single words found in *S. cerevisiae* (rows 1 and 2); single words conserved in three or more *Saccharomyces* genomes (rows 3 and 4); word pairs found in *S. cerevisiae* (row 5); word pairs conserved in three or more *Saccharomyces* genomes (row 6). Arrows above the columns indicate conditions under which gene groups sharing the conserved word pair template (row 6) were significantly associated with gene expression changes, at a *p*-value of 0.001 (K-S test after FDR correction for multiple testing).

were more significantly different from the average expression of genes in the genome when either conservation or word pairs was used as an additional criterion for gene selection. In Figure 3.8, the last two rows for each word pair indicate the average expression profiles for genes that shared both words in *S. cerevisiae,* as well as the average expression profile for genes that conserved both words in multiple genomes, respectively. Strikingly, the consideration of word pair conservation yielded further increases in average gene expression profiles compared to word pairs in *S. cerevisiae* alone. Thus, conserved word pair templates contained more specific predictors of gene expression than comparable sequence templates derived from *S. cerevisiae* alone.

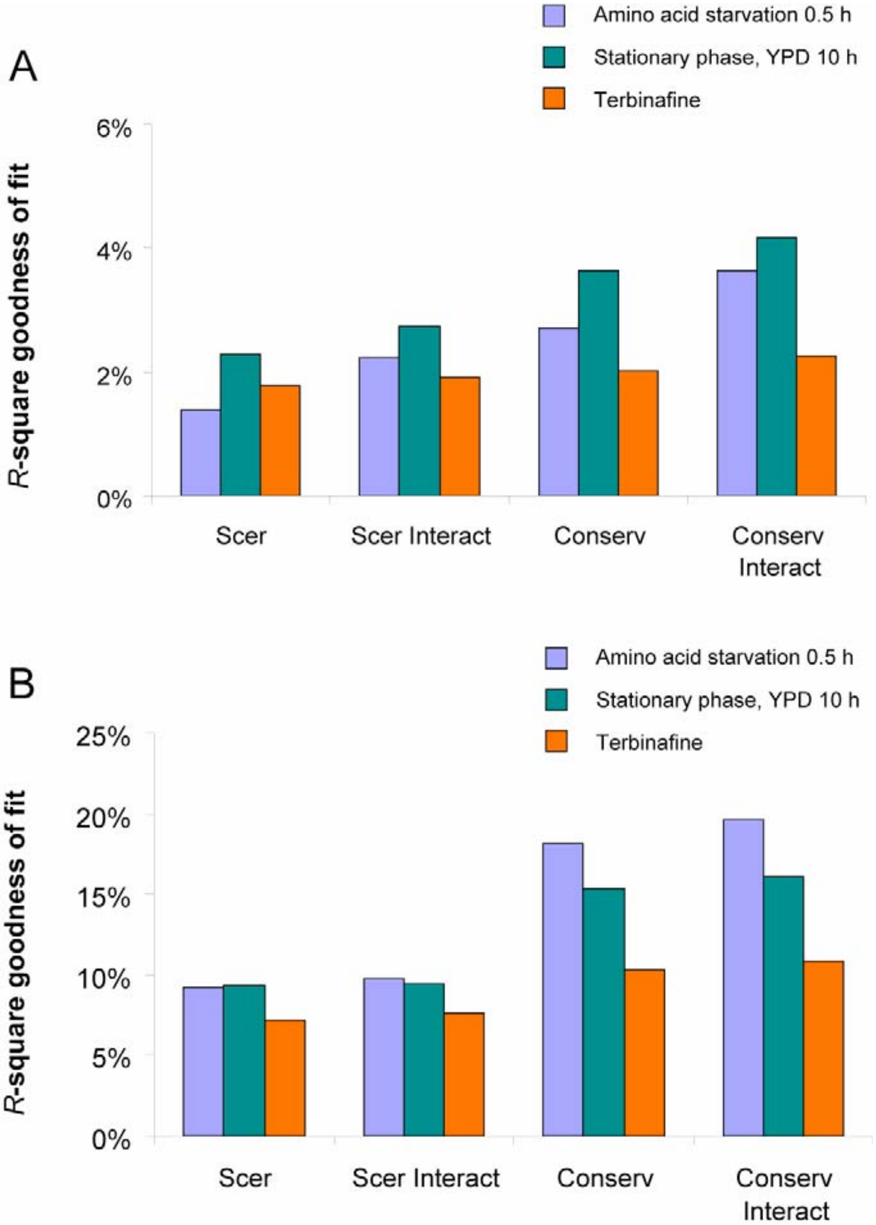
To evaluate how well sequence features can explain gene expression changes across the whole genome, several groups have constructed linear regression models using various choices for features (Bussemaker *et al.*, 2001; Keles *et al.*, 2002; Wang *et al.*, 2002; Conlon *et al.*, 2003). The *R*-square statistic of a regression model indicates the percent of global variance that can be explained using the sequence features in the model. Models with better fits to the genome-wide expression data would thus have greater*R*-square values. To assess the sensitivity of individual word pairs in explaining global gene expression, we first constructed regression models using individual word pairs (see Materials and Methods). We chose three representative environmental conditions: amino acid starvation (30 min) (Gasch *et al.*, 2000); stationary phase (10 h in YPD) (Gasch *et al.*, 2000); and ergosterol inhibition (terbinafine) (Hughes *et al.*, 2000b). We constructed regression models using counts of individual words in *S. cerevisiae* TCRs, or using counts of words that were conserved among *Saccharomyces* TCRs. Sequence conservation improved the fit of regression models based on individual word pairs

(ΔR^2 = 0.3% to 1.3%) (Figure 3.9A). These results clearly show that sequences conserved in multiple *Saccharomyces* species were more likely to be associated with gene expression changes.

To assess the joint contribution of word pairs on gene expression, we also included interaction terms between the individual words only if their coefficients were statistically significant (see Materials and Methods). Pairwise interaction terms, expressed as the product of scores for two features, represent a standard way to assess whether two features contribute non-additively to gene expression (Keles *et al.*, 2002). Indeed, the inclusion of significant pairwise interaction terms improved the fits for both the *S. cerevisiae* sequence model and the conserved sequence model, increasing the *R*-square by a further 0.2% to 0.9% (Figure 3.9A). Whereas the interaction terms only comprise a small proportion of the global variance, they can be interpreted as statistical evidence of dependence between sequence features (Keles *et al.*, 2002). Therefore, the non-additive contributions of conserved word pair templates further suggest their involvement in multifactorial regulation.

We expanded these models to include multiple conserved word pair templates using a stepwise linear regression procedure. The set of potential sequence features was expanded to include all words found in templates associated with significant gene expression changes in that condition, as assessed previously by the K-S test (see Materials and Methods). The final *R*-square values for regression models based on occurrences of multiple words in *S. cerevisiae* ranged from 7.2% to 9.3% (Figure 3.9B and Table 3.3). Once again, the use of conserved instances of individual words yielded better model fits, with improvements in *R*-square values from 3.1% to 9.5%.

Figure 3.9) Regression models using conserved word pairs represented better fits to genome-wide expression data



(A) Linear regression models were fit using two words found in a representative conserved word pair template, using different sequences features as predictors: counts of individual words found in the TCRs of *S. cerevisiae* genes (Scer); word counts in

Figure 3.9 (continued)

S. cerevisiae along with a pairwise interaction term (Scer Interact); counts of word that were conserved in the TCRs of three or more *Saccharomyces* genomes (Conserv); conserved word counts along with a pairwise interaction term (Conserv Interact). The sequence features and gene expression datasets used were: CACGTG and TGTGGC for 30 minutes after amino acid starvation (violet) Gasch *et al.*, 2000; ACCCCA and CCGCCG for 10 hours of growth in YPD (green) Gasch *et al.*, 2000; and CCGATA and TCGTTT for 3 hours of growth of terbinafine (orange) Hughes *et al.*, 2000b. The chi-square goodness-of-fit values obtained from the best models is summarized for each model.

(B) Same as above, expect that all words found in conserved word pair templates were used as possible features for stepwise linear regression models.

Further improvements in the model fit (ΔR^2 = 0.5% to 1.4%) were obtained using pairwise interaction terms between individual words found in the same word pair template. The total *R*-square values for the regression models based on conserved word pair templates with interaction terms thus ranged from 10.8% to 20.2% (Table 3.3). Thus, sequences features based on conserved word pair templates could explain more of the global gene expression changes than features based on *S. cerevisiae* alone.

Table 3.3) Stepwise linear regression statistics

(A) Amino Acid Starvation 0.5 h

	S. cerevisiae			Three or more genomes			
Word	$\boldsymbol{M_f}$	*p*-val	Δ R^2	$\boldsymbol{M_{fc}}$	*p*-val	Δ R^2	Wang *p*-val
AAATTT	-0.165	< 2.0e-16	3.1%	-0.293	< 2.0e-16	6.6%	1.3e-37
GATGAG	—	—	—	-0.333	6.7e-16	4.1%	1.6e-30
AAGGGG	0.209	3.2e-14	1.8%	0.455	< 2.0e-16	3.5%	1.6e-22
TGTGGC	0.094	1.1e-03	0.6%	0.283	2.9e-07	1.6%	5.0e-07
CCCTTA	0.300	**2.0e-16**	1.7%	0.363	**< 2.0e-16**	1.4%	3.2e-06
TGACTC	0.229	**4.6e-10**	0.8%	0.311	**2.2e-11**	1.0%	4.6e-01
AAATTT* GATGAG	—	—	—	-0.266	**9.1e-09**	0.5%	—
CACGTG	0.045	3.5e-01	0.5%	0.146	8.6e-03	0.5%	1.9e-07
CACGTG* TGTGGC	0.443	**3.8e-10**	0.5%	0.749	**1.0e-12**	0.9%	—
GTGAAA	-0.066	**1.1e-03**	0.3%	-0.082	**4.6e-03**	0.1%	—
TCTTTT	-0.022	2.3e-02	0.1%	—	—	—	—
Total R^2			9.6%			20.2%	

(B) Stationary phase, YPD 10 h

	S. cerevisiae			Three or more genomes			
Word	$\boldsymbol{M_f}$	*p*-val	Δ R^2	$\boldsymbol{M_{fc}}$	*p*-val	Δ R^2	Wang *p*-val
AAATTT	-0.218	< 2.0e-16	3.2%	-0.377	< 2.0e-16	5.8%	5.5e-39
AAGGGG	0.233	1.7e-11	0.9%	0.591	< 2.0e-16	4.0%	4.5e-26
CCCTTA	0.460	**< 2.0e-16**	3.7%	0.579	**< 2.0e-16**	2.2%	3.0e-07
GATGAG	—	—	—	-0.242	4.1e-06	1.8%	4.4e-18
ACCCCA	0.224	**3.0e-03**	0.3%	0.459	**1.5e-06**	1.0%	—
AAATTT* GATGAG	—	—	—	-0.287	**1.7e-06**	0.4%	—
CCGCCG	0.333	**5.1e-07**	0.8%	0.208	**1.5e-02**	0.3%	—
ACCCCA* CCGCCG	0.294	**1.8e-02**	0.1%	0.807	**5.6e-05**	0.3%	—
GTGAAA	-0.090	**4.2e-04**	0.2%	-0.122	**1.0e-03**	0.2%	—
Total R^2			9.4%			6.0%	

Table 3.3 (continued)

(C) Terbinafine 3 h

	S. cerevisiae			Three or more genomes			
Word	M_f	*p*-val	Δ R^2	M_{fc}	*p*-val	Δ R^2	Wang *p*-val
TGACTC	0.162	**< 2.0e-16**	3.5%	0.261	**< 2.0e-16**	5.1%	1.3e-14
TCGTTT	0.071	< 2.0e-16	2.0%	0.132	< 2.0e-16	3.3%	2.5e-24
TGAAAC	-0.055	**1.3e-12**	1.1%	-0.077	**9.50e-11**	0.9%	4.0e-03
GATGAG	-0.029	**1.7e-03**	0.3%	-0.047	**6.70e-06**	0.4%	—
AAGGGG	0.025	**1.1e-02**	0.1%	0.050	**5.40e-04**	0.3%	2.4e-01
CCGATA	-0.008	6.5e-01	0.1%	0.004	8.6e-01	0.1%	—
CCGATA* TCGTTT	0.080	**9.4e-06**	0.3%	0.146	**3.2e-07**	0.5%	—
CCCTTA	-0.021	**5.0e-02**	0.1%	-0.038	**1.1e-02**	0.1%	—
Total R^2			7.6%			10.8%	

Words and pairwise interaction terms are reported in the order of selection by the stepwise linear regression procedure performed on conserved words. The influence terms (M_f), associated *p*-values, and increase in *R*-square values were computed using the statistical package R. Wang *et al.*, 2002 previously fit regression models using sequence features derived from *S. cerevisiae*. The *p*-values of the most similar sequences features in their regression models were reported where available; sequence features that were more significant in this analysis are indicated in bold.

DISCUSSION

This work describes two principles for analyzing combinations of regulatory sequences. First, sequence conservation among closely related yeast species was used to find sequences that were more likely to be functionally important. Secondly, a template approach that considered joint positional distributions of word pairs increased the specificity of gene expression predictions using sequence-based rules. We have demonstrated that higher-order sequence features within TCRs were conserved across multiple *Saccharomyces* genomes. Closely spaced and jointly conserved word pairs were also more likely to be associated with specific gene expression changes. A large proportion of words contained in templates matched known transcription factor binding sites. In many cases, associations between templates and gene expression changes were significant in conditions when the corresponding transcription factors are known to be active. In addition, groups of genes that co-conserved both words in a template often were enriched for common functional roles. These results suggest that conserved word pair templates, which were discovered strictly based on higher-order properties of sequence conservation, also carry biological relevance.

Conserved word pair templates may be consistent with several underlying biochemical mechanisms. One possible interpretation of templates is that closely spaced sequence pairs may promote direct or indirect interactions between transcription factors by increasing the local concentrations of the individual factors. For example, the proximity of Cbf1 and Met31 or Met32 binding sites may promote interaction between these factors in recruiting their common transcriptional activators, Met4 and Met28. Experimental studies on the TCRs of *MET3* and *MET28* have demonstrated that the binding of Cbf1 enhances the DNA binding affinity of Met31 or Met32 (Blaiseau and Thomas, 1998). Indeed, biochemical experiments suggest that all of these proteins may interact at the TCRs of some sulfur utilization genes (Blaiseau and Thomas, 1998).

Another possible regulatory scheme for conserved, closely-spaced word pairs is that individual sequences found in templates may correspond to binding sites for transcription factors that bind independently under the same or separate conditions. The Msn2/4 and Hsf1 transcription factors, whose binding sites were similar to words identified in a template, represent an example of multifactorial regulation in response to distinct environmental stimuli (Gasch, 2003). Spacing constraints between their binding sites could nevertheless be important under conditions when both factors are active. Recent experiments have suggested that transcription factors that do not interact may still coactivate gene expression as long as their binding sites are spaced within a nucleosome length (~150 bp), due to collaborative competition of the bound transcription factors with core histones (Miller and Widom, 2003).

Close spacing between word pairs may be important for reasons other than the promotion of transcription factor interactions. Different regions of TCRs at varying windows away from translation start sites may be more competent at recruiting or inhibiting RNA polymerase. These differences may be influenced by nucleosome accessibility, chromatin structure, or DNA physical properties, which can be correlated with local A+T content (see Liao *et al.*, 2000 for references). Notably, we have also found that the relative proportions of A and T nucleotides vary considerably within the 200 bp closest to translation start sites (A. M. Moses, M. B. Eisen and Audrey Gasch, unpublished results). Low-complexity words that contained 4 or more A's or T's could be found in many templates (denoted by T_nC in Figure 3.6 and Table 3.1); these words may serve as surrogates for a distance window from translation start. Binding sites that are closely spaced to these low-complexity words may be found in more transcriptionally competent regions of TCRs. Alternatively, the possibility that each word in an identified pair may be found at similar distances from a third conserved sequence element in all TCRs cannot be discounted.

Direct biological models of binding site organization in TCRs, as exemplified by conserved word pair templates, provided several advantages over statistical models based on sequence combinations in *S. cerevisiae*. Average gene expression profiles showed that conserved word pairs were more specific predictors of gene expression (*i.e.*, much fewer false positives) than single or pairwise sequences derived from *S. cerevisiae*, indicating that conserved regions among these closely related *Saccharomyces* species were enriched for functional sequences. The consideration of distance constraints between pairs of conserved sequences found many more examples than a previous study of binding site clustering for multiple transcription factors in *S. cerevisiae* (Wagner, 1999). In addition, we discovered new sequences and pairwise interaction terms using regression models similar to those reported in Wang *et al.*, 2002 and Conlon *et al.*, 2003 (Table 3.3). Conserved word pair templates accounted for similar changes in genome-wide expression (*R*-square from ~11% to ~20%) using only 8 to 10 features, compared with dozens of overlapping features used by other methods (Wang *et al.*, 2002; Conlon *et al.*, 2003). Therefore, individual features from our methods were more predictive about genome-wide expression changes.

A key limitation of our approach is the use of hexamers, which may fail to capture known binding sites. For example, the binding sites for Mcm1 and Rap1 are poorly modeled by exact words, in that these transcription factors bind sequences with relaxed specificity at certain positions (Stormo, 2000). Our method missed examples of multifactorial regulation involving Mcm1 or Rap1 that were suggested by previous work using position weight matrices (Pilpel *et al.*, 2001). In addition, our method required sequence identity for a word to be labeled as conserved. This strict requirement omitted binding sites that may have retained their function, despite mutations in degenerate positions that may have little impact on transcription factor binding. This tradeoff between enumerating all possible words and capturing degenerate positions in binding sites was compounded by the very large number of pairwise word combinations that were enumerated. Further work should incorporate more complicated sequence models, as well as optimization methods that restrict the search space of sequence combinations.

The consideration of joint conservation and close spacing has provided insights into how TCR organization may influence the multifactorial regulation of gene expression in *Saccharomyces cerevisiae*. These criteria were motivated by experimental studies on the positional organization of individual binding sites within TCRs, with the hypothesis that this underlying architecture would be functionally conserved. Even more complicated higher-order sequence rules are apparent in the organization of *cis*-regulatory modules in *Drosophila melanogaster* (Berman *et al.*, 2002). Nevertheless, a common organizational theme of the TCRs in both of these organisms is the importance of relative spacing between transcription factor binding sites. The discovery of additional principles for TCR organization will further advance our understanding of how regulatory information is encoded in genome sequences.

ACKNOWLEDGEMENTS

We are indebted to Audrey Gasch for her insightful advice and suggestions. We thank Justin Fay for a critical reading of the manuscript and help with the permutation test; Peter Bickel and John Storey for their statistical advice; and Nabuo Ogawa and Patrick Brown for sharing gene expression data on cadmium conditions prior to publication. D.Y.C. is a Howard Hughes Medical Institute Predoctoral Fellow, and M.B.E. is a Pew Scholar in the Biomedical Sciences. This work was conducted under the US Department of Energy contract No. ED-AC03-76SF00098.

CHAPTER 4

PROMOTER ARCHITECTURE OF

YEAST SULFUR UTILIZATION GENES

PREFACE

The last two chapters have provided genome-wide insights on the association of regulatory sequences with changes in gene expression. In particular, the previous chapter presented many examples of multifactorial regulation that were inferred from close spacing between conserved word pairs. These distance constraints probably differ for pairs of transcription factors that interact by different mechanisms. By focusing experiments on a model pair of yeast transcription factors, I have characterized the effects of distance constraints and sequence contexts on recruitment of the co-activator, Met4.

ABSTRACT

Organizational features of regulatory sequences, such as the distance and sequence composition between transcription factor binding sites, influence the assembly of the multiprotein complexes that regulate RNA polymerase recruitment. We expect that different constraints on promoter architecture may arise from distinct mechanisms of transcription factor interactions at regulatory regions. We have developed a genetic approach to investigate how reporter gene transcription is affected by varying the spacing between binding sites for transcription factors known to coordinately regulate transcription. We characterized the components of promoter architecture that govern the yeast transcription factors Cbf1 and Met31 or Met32, which bind independently, but collaboratively recruit the co-activator Met4. A Cbf1 binding site was required upstream of a Met31 or Met32 binding site for full reporter gene activation. Distance constraints on coactivator recruitment were more flexible than those for cooperatively binding transcription factors. Distances from 18 to 50 bp between binding sites could support efficient recruitment of Met4, with only slight modulation by helical phasing. Intriguingly, we found that certain sequence contexts between the binding sites abolished gene activation. These results yield insight into the influence of both binding site architecture and local DNA flexibility on gene activation, and can be used to refine computational predictions of gene expression from promoter sequences.

BACKGROUND

The design principles of gene regulation are intricately complicated. Differential expression of an organism's genetic repertoire alters its phenotypic response to varying environmental conditions. Just a few nucleotide changes in the regulatory region of a single gene can affect its expression level (Bond *et al.*, 2004; Liao *et al.*, 2004; Rockman *et al.*, 2004). By contrast, other genes are regulated more robustly and can tolerate multiple changes in their regulatory regions while preserving gene expression output (Ludwig *et al.*, 2000; Wray *et al.*, 2003). The understanding of how sequence information determines transcriptional regulation should enable the design of highly sophisticated regulatory sequences that generate precise responses to specific conditions (Blackwood and Kadonaga, 1998; Davidson *et al.*, 2002; Guet *et al.*, 2002; Setty *et al.*, 2003; McAdams *et al.*, 2004). Yet to fully exploit the combinatorial logic inherent to gene regulation, we need more mechanistic details of how multiple binding sites are integrated within a regulatory region.

In most eukaryotes, multiple transcription factors interact at transcriptional control regions to modulate levels of gene expression. Each transcription factor is activated by cellular signals in response to changes in environmental conditions. When bound to DNA, some transcription factor activators can anchor the assembly of multiprotein complexes that influence the recruitment of RNA polymerase. Efficient assembly often depends on the formation of optimally spaced protein-protein interactions among transcription factors and auxiliary proteins (Merika and Thanos, 2001; Ogata *et al.*, 2003; Remenyi *et al.*, 2004). Since transcription factors recognize specific sites on DNA, the distance between these binding sites can influence how transcription factors interact at regulatory regions. Overlapping sites may occlude two transcription factors from binding simultaneously, whereas sites spaced far apart require DNA looping for interactions to occur. Thus, the precise spatial arrangement of transcription factors in regulatory regions influences the level of gene activation. We use the term promoter architecture to refer both to distance constraints and to sequence context effects that govern interactions among transcription factor binding sites.

Several mechanisms that govern transcription factor interactions have been previously described. Transcription factors may bind cooperatively to adjacent sites in DNA, thus increasing the stability of the ternary DNA-protein complex. Since this effect is mediated by direct protein-protein interactions, sites for cooperatively binding transcription factors are usually spaced within 20 bp of each other (*e.g.*, Amin *et al.*, 1994; Hanes *et al.*, 1994; Brazas *et al.*, 1995; Drazinic *et al.*, 1996; Boros *et al.*, 2003Boros *et al.*, 2003). Alterations in spacing between the binding sites can drastically reduce gene activation unless helical phasing is preserved. Computational analyses suggest that helical phasing between predicted binding sites may be a general property of transcriptional control regions (Ioshikhes *et al.*, 1999; Makeev *et al.*, 2003).

Alternatively, transcription factors may bind to DNA independently and cooperatively recruit a coactivator protein. Co-recruitment of such activators is analogous to an “AND gate” in logic. Coincident binding of two proteins increases the fidelity and specificity of signal detection (Merika and Thanos, 2001; Naar *et al.*, 2001; Spiegelman and Heinrich, 2004). The network of transcription factors that regulates sulfur gene derepression in yeast provides a model system to dissect the promoter architecture requirements for coactivator recruitment. Among these transcription factors, only the coactivator, Met4, contains an activation domain. However, Met4 does not bind to DNA directly, but is recruited under sulfur limitation conditions by Cbf1 and Met28 to the *MET16* promoter, as well as by Met28 and Met31 or Met32 on regions from the *MET3* and *MET28* promoters (Kuras *et al.*, 1997; Blaiseau and Thomas, 1998). In addition, two-hybrid studies with Met4 truncation mutants revealed distinct regions that mediate interaction with Cbf1 and Met31 or Met32. Taken together, these studies suggest a model by which the co-activator Met4 is coordinately recruited by the transcription factors Cbf1, Met28 and Met31 or Met32 to the promoters of sulfur utilization genes. However, the effects of distance constraints and sequence context between Cbf1 and Met31 or Met32 binding sites have not been characterized.

Our goal is to understand how the constraints on promoter architecture differ for transcription factors that participate in coactivator recruitment, versus those that bind cooperatively. In this work, we developed a synthetic promoter assay to characterize how various distances between Cbf1 and Met31 or Met32 binding sites influenced gene activation in response to methionine starvation. The relative order of binding sites affected reporter gene activation. We discovered that distance constraints on coactivator recruitment were more flexible than those for cooperatively binding transcription factors. Distances from 18 to 50 bp between binding sites could support efficient recruitment of Met4, with only slight modulation by helical phasing. Intriguingly, we found that certain sequence contexts between the binding sites abolished gene activation. We noted that the probability of coactivator recruitment could be affected by the bendability of the spacer sequence between transcription factor binding sites.

MATERIALS AND METHODS

Plasmid construction

Plasmid pDC204 was constructed in five steps. 1) The *HIS3* coding region was PCR amplified from *S. cerevisiae* genomic DNA using the primers HIS3_F_BamHI and HIS3_R (Table 4.1) and cloned downstream of the *MEL1* minimal promoter (P_{MEL1}) by ligating into the BamHI + EcoRV-cleaved plasmid YIpMELβ2 from EUROSCARF (Melcher *et al.*, 2000). Two changes were then made to the *MEL1* minimal promoter. 2) An NcoI site was introduced into P_{MEL1} 31 bp upstream of the existing XhoI site by site-directed mutagenesis (oligos MEL1_NcoI_W and MEL1_NcoI_C). 3) An out-of-frame ATG codon located 17 bp upstream of the *HIS3* coding region was removed by site-directed mutagenesis (oligos ATG_W and ATG_C). 4) The P_{MEL1}-*HIS3* fusion construct was PCR amplified (primers pMH14-F_ApaI & pMH14-R_AscI-SacII) and cloned into the ApaI + SacII-cleaved plasmid pRS314 Sikorski and Hieter, 1989. 5) The *Kluyveromyces lactis LEU2* gene was PCR amplified from pUG73 (primers pUG73_F and pUG73_R) (Gueldener *et al.*, 2002) and cloned into the AscI site of the above plasmid. Restriction digests confirmed the same-strand orientation of the HIS3 and LEU2 coding regions, and sequencing verified the promoter and coding regions.

Promoter library construction

Degenerate oligonucleotides were designed with a Cbf1 binding site at a fixed distance upstream of a Met31 or Met32 binding site (Operon) (Table 4.1). Ten bp of flanking sequence upstream of the Cbf1 binding site and downstream of the Met31 or Met32 binding site were included from the wild-type *MET16* promoter. Double-stranded DNA was

Table 4.1) List of oligonucleotides used in this study

Plasmid construction

Oligo name	Sequence (5’→3’) [restriction sites underlined]
HIS3_F_BamHI	CGGGATCCCGAAGATGACAGAGCAGAAAGC
HIS3_R	CCTCGTTCAGAATGACACG
MEL1_NcoI_W	CCCTGAAAGGTTTTTCCATGGAATAGTCAGGACGC
MEL1_NcoI_C	GCGTCCTGACTATTCCATGGAAAAACCTTTCAGGG
ATG_W	GTAATAAAAGCAACGACGTTGAACGGATCCCGAAAG
ATG_C	CTTTCGGGATCCGTTCAACGTCGTTGCTTTTATTAC
pMH14-F_ApaI	ATAGGGCCCGGAAATTTGTGTAAAACCCCC
pMH14-R_AscI-SacII	AACAACCGCGGATAATGGCGCGCCCCTCGTTCAGAATGACAC
pUG73_F	AAGGCGCGCCGCATAGGCCACTAGTGGATCTG
pUG73_R	AGTAAGGCGCGCCCAGCTGAAGCTTCGTACGC

Promoter library construction

Oligo name	Sequence (5’→3’) [restriction sites underlined]
MET16_reverse	CCGCTCGAGTTACTGAAGTTG
Cbf1_6_Met31	CATGCCATGGTATCATCATTTCACGTGGCCACAACTTCAGTAACTCGAGCGG
Cbf1_8_Met31	CATGCCATGGTATCATCATTTCACGTGGNNCCACAACTTCAGTAACTCGAGCGG
Cbf1_10_Met31	CATGCCATGGTATCATCATTTCACGTGGNNNNCCACAACTTCAGTAACTCGAGCGG
Cbf1_12_Met31	CATGCCATGGTATCATCATTTCACGTGGNNNNNNCCACAACTTCAGTAACTCGAGCGG
Cbf1_14_Met31	CATGCCATGGTATCATCATTTCACGTGGNNNNNNNNCCACAACTTCAGTAACTCGAGCGG
Cbf1_16_Met31	CATGCCATGGTATCATCATTTCACGTGGNNNNNNNNNNCCACAACTTCAGTAACTCGAGCGG
Cbf1_18_Met31	CATGCCATGGTATCATCATTTCACGTGGNNNNNNNNNNNNCCACAACTTCAGTAACTCGAGCGG
Cbf1_20_Met31	CATGCCATGGTATCATCATTTCACGTGGNNNNNNNNNNNNNNCCACAACTTCAGTAACTCGAGCGG
Cbf1_22_Met31	CATGCCATGGTATCATCATTTCACGTGGN_{18}CCACAACTTCAGTAACTCGAGCGG
Cbf1_24_Met31	CATGCCATGGTATCATCATTTCACGTGGN_{20}CCACAACTTCAGTAACTCGAGCGG

Cbf1_26_Met31	CATGCCATGGTATCATCATTTCACGTGGN_{22}CCACAACTTCAGTAACTCGAGCGG
Cbf1_28_Met31	CATGCCATGGTATCATCATTTCACGTGGN_{24}CCACAACTTCAGTAACTCGAGCGG
Cbf1_30_Met31	CATGCCATGGTATCATCATTTCACGTGGN_{26}CCACAACTTCAGTAACTCGAGCGG
Cbf1_32_Met31	CATGCCATGGTATCATCATTTCACGTGGN_{28}CCACAACTTCAGTAACTCGAGCGG
Cbf1_34_Met31	CATGCCATGGTATCATCATTTCACGTGGN_{30}CCACAACTTCAGTAACTCGAGCGG

synthesized by Bio-X-Act polymerase (Bioline) from the primer MET16_reverse (Table 4.1), digested with NcoI and XhoI and ligated into pDC204.

Yeast strains and media

Strain BY4742 (*MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0*) was obtained from Invitrogen. Plasmids were introduced into this parental strain by lithium acetate transformation (Gietz and Woods, 2002) and subsequent selection on dropout media lacing leucine. To measure growth rates in liquid culture, over 100 yeast transformants containing plasmids of the same promoter size were pooled and cultured overnight in dropout media lacking leucine. To induce reporter gene expression, yeast cultures were diluted to early log phase (OD_{600} ~ 0.04) in dropout media lacking leucine, histidine and methionine, then grown at 30°C with shaking at 250 rpm. After 3 hours, each culture was split in half and 3-aminotriazole (Sigma) was added to one culture. Timepoints were commenced 2.5 hours after 3-aminotriazole addition.

Association of promoter architectures with gene expression changes

Gene expression data for sulfur-limiting conditions was obtained from the 4 replicates of cadmium treatment reported by Fauchon *et al.*, 2002, as well as the first 4 timepoints of the amino acid starvation timecourse from Gasch *et al.*, 2000. An activation level was summarized for each gene by averaging the log base 2 expression ratios. Matches to Cbf1 (TCACGTG) and Met31 or Met32 (TGTGGC) consensus sequences were scored for each promoter, which was defined as the 500 bp upstream of the translation start site. For various binding site combinations, the average activation level was calculated for genes that shared the appropriate combinations of binding sites.

RESULTS

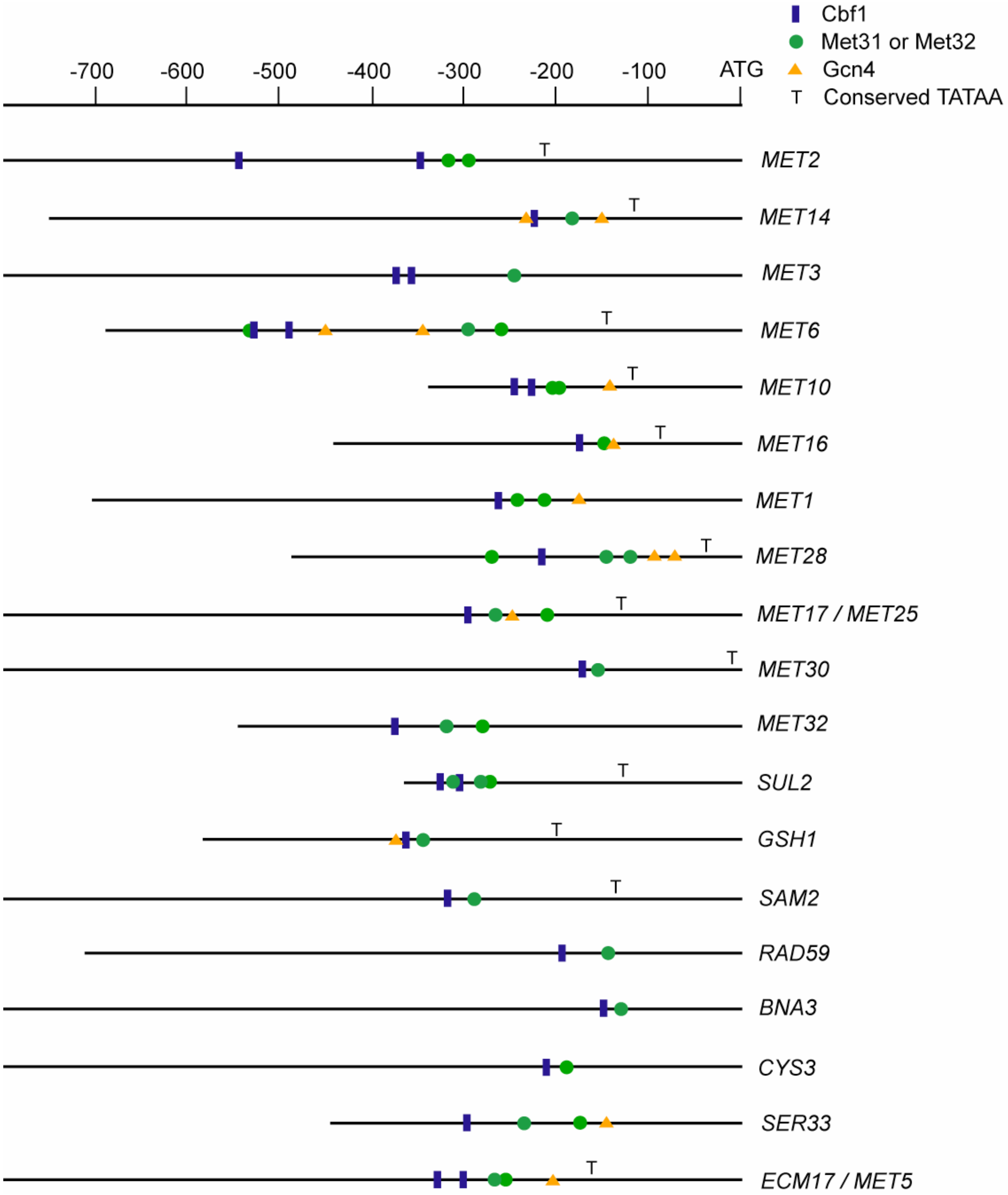
Annotated promoters of sulfur-regulated genes contain closely spaced binding sites

We examined the promoter architectures of 20 genes in *S. cerevisiae* that are annoated as being co-expressed under sulfur-limiting conditions (Thomas *et al.*, 1992; Balakrishnan *et al.*, 2005). All of these promoters contained Cbf1 and Met31 or Met32 binding sites that were perfectly conserved among at least 3 of 4 closely-related yeast species (Cliften *et al.*, 2003; Kellis *et al.*, 2003) (Figure 4.1). Each sulfur-regulated promoter included a Cbf1 binding site upstream of a Met31 or Met32 binding site. A histogram of distances between the closest pair of Cbf1 and Met31 or Met32 binding sites showed a peak between 10 and 30 bp (Figure 4.2A). This peak suggested an optimal distance between the transcription factors was necessary for efficient Met4 recruitment. When investigating whether the distances between the closest pairs of binding site were helically phased, we could not detect a significant enrichment of distances on a certain face of DNA (Figure 4.2B). Finally, the vast majority of annotated promoters contained Met31 or Met32 binding sites within 100 to 350 bp upstream of the translation start site (Figure 4.2C). Thus, computational sequence analyses suggests that distance constraints are important for the recruitment of the coactivator, Met4.

Cbf1 binding sites upstream of Met31 or Met32 binding sites yielded maximal activation

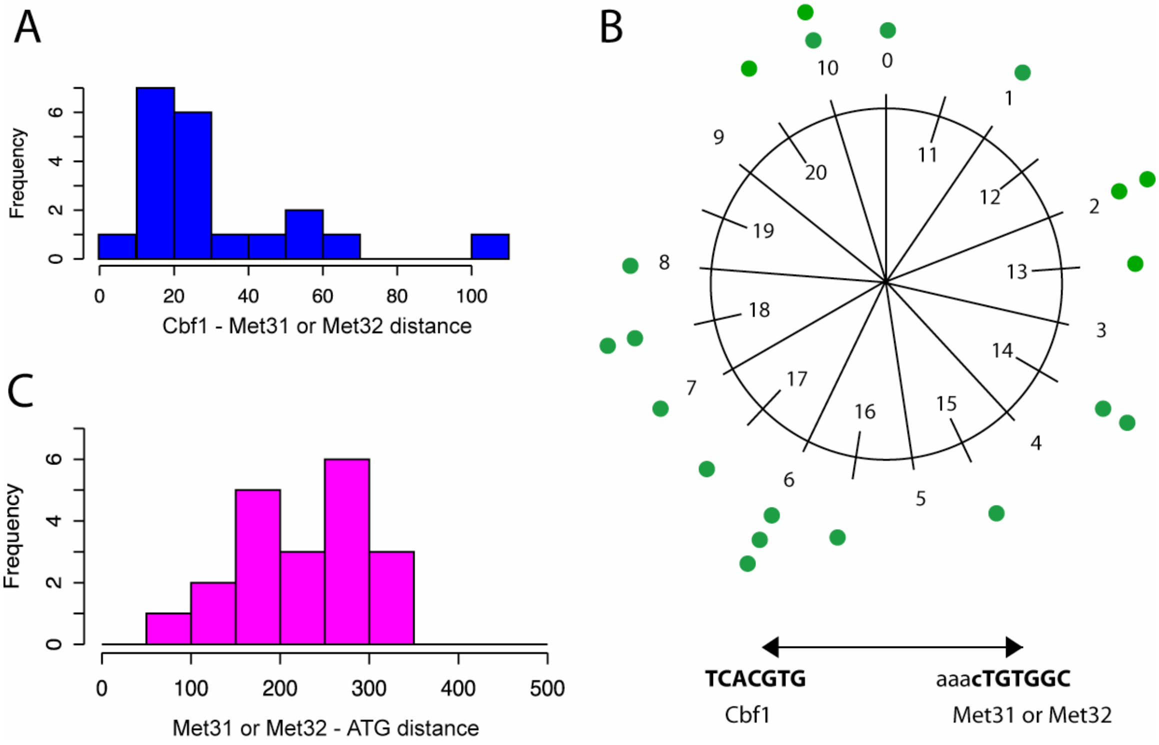
A larger collection of sulfur-regulated promoters would provide more statistical power to define key components of promoter architecture. To explore a sequence space more diverse than that found in the yeast genome, we developed a synthetic genetic approach to select for sulfur-regulated promoters from a plasmid library (Figure 4.3).

Figure 4.1) Promoter architectures of annotated sulfur-regulated genes



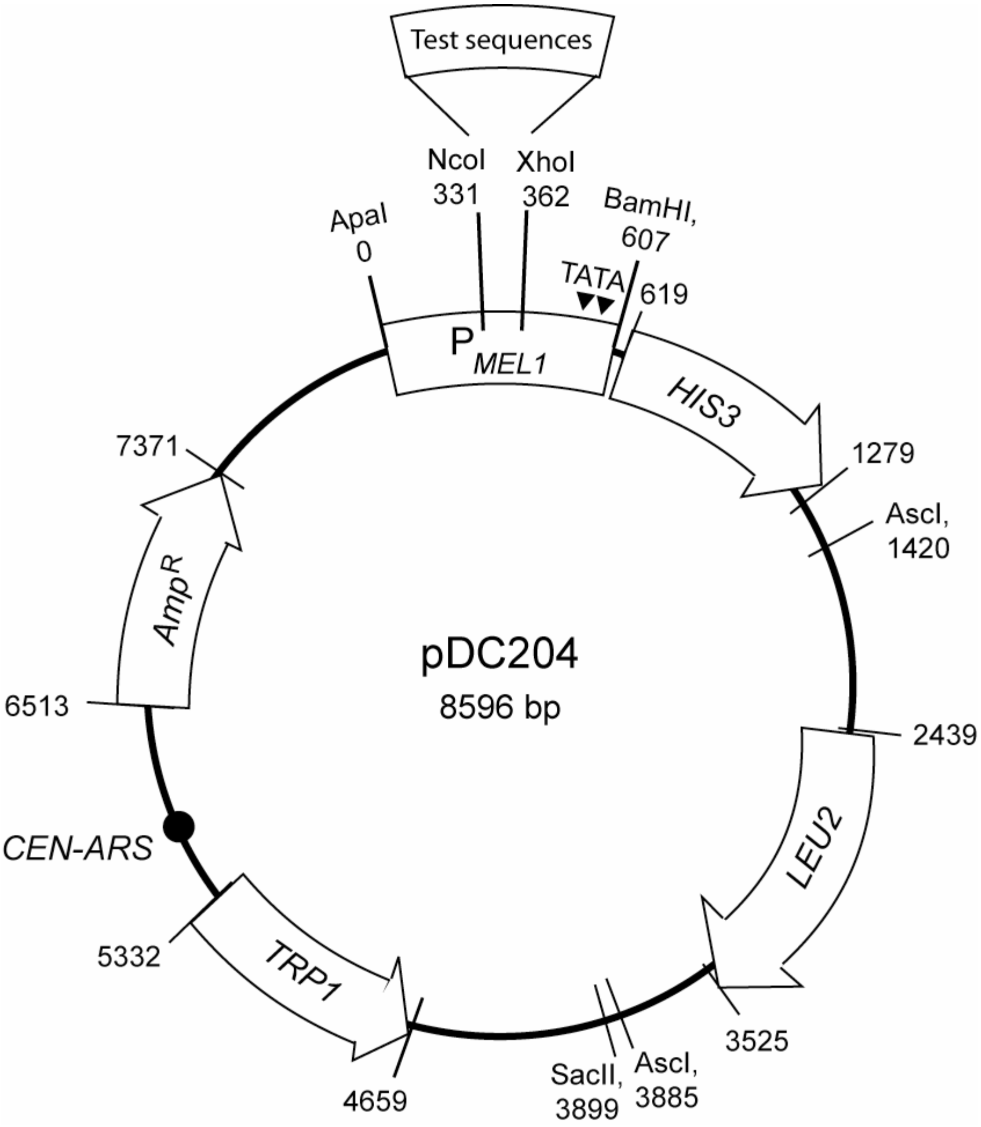
Conserved binding sites for Cbf1 (blue rectangles, TCACGTG), Met31 or Met32 (green circles, TGTGGC), Gcn4 (orangle triangles, TGA[C|G]TCA) and TBP (TATAA) are drawn to scale in the indicated intergenic regions. A binding site was considered conserved if at least 3 invariantly copies were aligned in a multiple sequence alignment of closely-related *Saccharomyces* species (Cliften *et al.*, 2003; Kellis *et al.*, 2003).

Figure 4.2) Minimum distances between conserved Cbf1 and Met31 or Met32 binding sites in annotated sulfur-regulated promoters



(A) Histogram of minimum distances between a Cbf1 binding site (TCACGTG) and a Met31 or Met32 (TGTGGC) binding site. Distances were calculated from the center of each binding site, as indicated by the arrows between the consensus sequences. (B) Helical wheel projection of minimum distances. Cbf1 binding sites were aligned at the top of the helical wheel (position 0). Each green circle represents the remainder of a minimum distance from (A) divided by 10.5 bp. Since the helical pitch of DNA is 10.4 bp, each circle approximates the position of the Met31 or Met32 binding site relative to the Cbf1 binding site. (C) Histogram of distances between the Met31 or Met32 binding sites from (A) and the translation start site.

Figure 4.3) Synthetic promoter system



A minimal promoter from the *MEL1* gene (*P_{MEL1}*) was fused upstream of a *HIS3* reporter gene on a single-copy plasmid. Selected restriction enzyme sites are labeled with their coordinates. Various combinations of Cbf1 and Met31 or Met32 binding sites were inserted between the NcoI and XhoI restriction enzymes sites in the promoter.

We engineered a single-copy plasmid that fused a minimal promoter upstream of the *HIS3* reporter gene. To test their transcriptional activation potential, different promoter architectures were embedded in the context of the minimal promoter from the *S. cerevisiae MEL1* gene. This promoter was chosen for its low background expression, compared to promoters derived from the *S. cerevisiae CYC1* gene (Melcher *et al.*, 2000). Promoter architectures with combinations of regulatory sequences that were sufficient to induce expression of the *HIS3* reporter gene enabled the parental yeast strain BY4742 to grow in media lacking histidine. In addition, semiquantitative measurements of *HIS3* expression can be assayed by titration with 3-amino-1,2,4-triazole (3-AT), a competitive inhibitor of the His3 gene product (Horecka and Sprague, 2000). Faster growth rates in the presence of larger concentrations of 3-AT correspond to higher expression levels of the *HIS3* gene.

We sought to define the minimal regulatory information that was sufficient to induce reporter gene expression in the absence of methionine. Neither the minimal promoter alone nor a single Met31 or Met32 binding site could induce *HIS3* expression in inducing conditions with 10 mM 3-AT (Figure 4.4A). A single Cbf1 binding site supported weak growth on 10 mM 3-AT. In the wild-type *MET14* promoter, a Cbf1 binding site was found 35 bp upstream from a Met31 or Met32 binding site, as measured by center-to-center distance. Two Cbf1 binding sites placed at the same distance showed moderate *HIS3* expression. In contrast, two Met31 or Met32 binding sites were unable to support growth. A promoter with a Cbf1 binding site upstream of a Met31 or Met32 binding site showed the highest level of *HIS3* expression, whereas a promoter with the reverse order of binding sites was unable to support growth on 10 mM 3-AT.

Figure 4.4) Reporter gene expression driven by various combinations of Cbf1 and Met31 or Met32 binding sites matches computational predictions

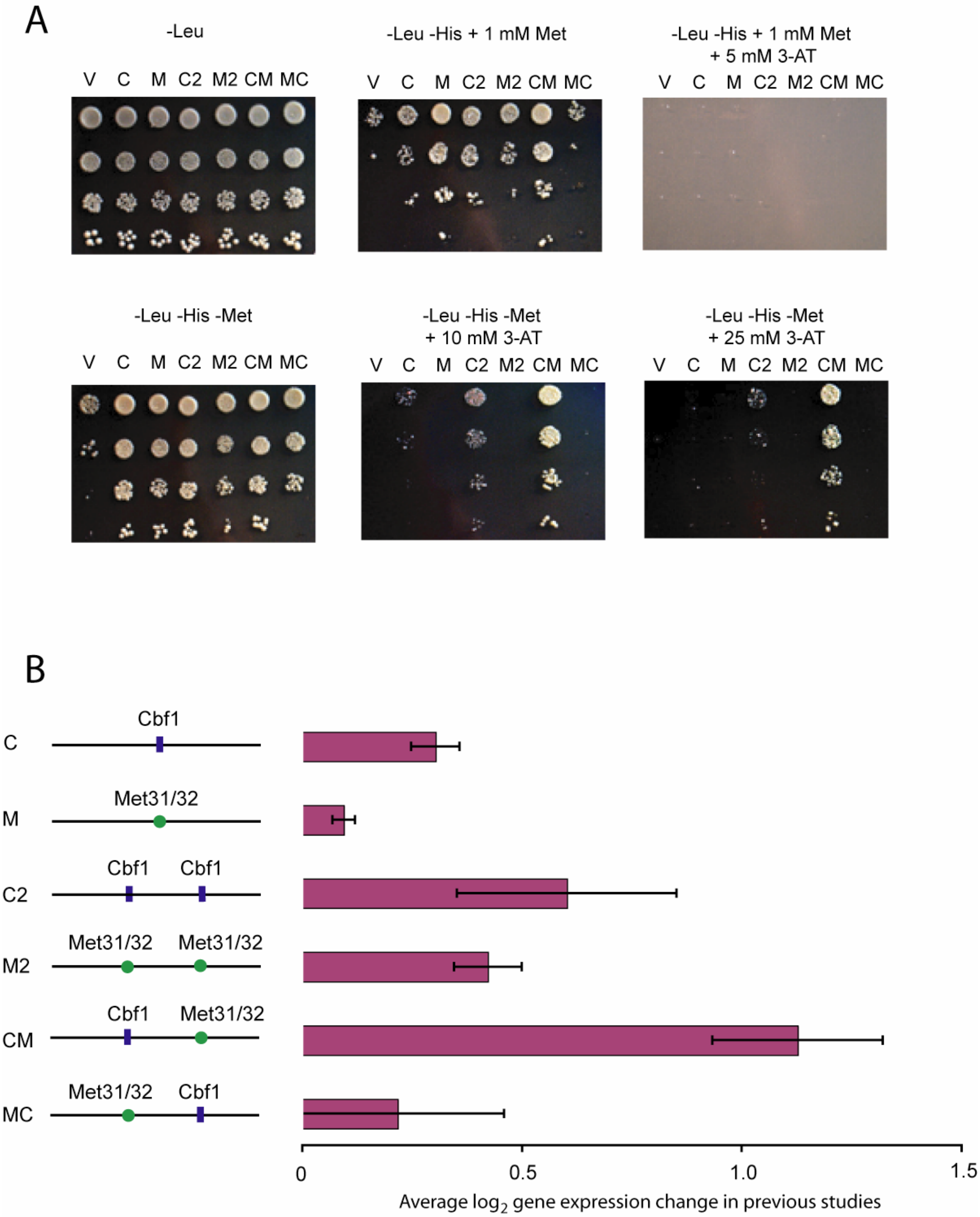


Figure 4.4 (continued)

(A) Serial dilutions of yeast containing a reporter plasmid with a different binding site combination, labeled as follows. V: vector alone, C: Cbf1 binding site, M: Met31 or Met32 binding site, C2: Two Cbf1 binding sites spaced by 35 bp, M2: Two Met31 or Met32 binding sites spaced by 35 bp, CM: Cbf1 binding site placed 35 bp upstream of a Met31 or Met32 binding site, MC: Met31 or Met32 binding site placed 35 bp upstream of a Cbf1 binding site. Yeast strains were grown on the indicated media for 5 days at 30°C. (B) Average gene induction in sulfur-limitation conditions associated with endogenous genes containing various promoter architectures. An induction level for each gene was calculated as the average log base 2 expression ratio from previously published gene expression studies (Gasch *et al.*, 2000; Fauchon *et al.*, 2002). Induction levels were averaged over sets of genes that shared the indicated binding site combinations in the 500 bp upstream of their translation start sites.

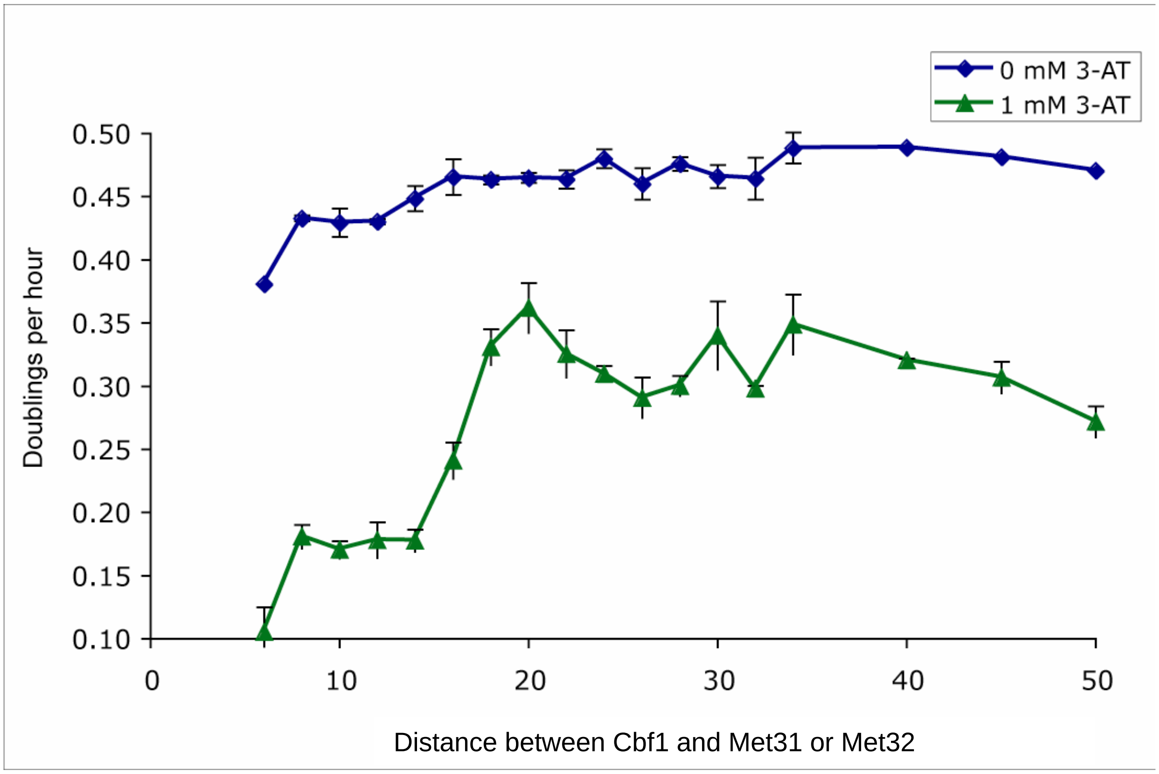
These results confirmed computational predictions of the activation potentials for the various promoter architectures (Figure 4.4B). Using published microarray studies of sulfur limitation conditions, we calculated the average induction of genes whose promoters shared combinations of Cbf1 or Met31 or Met32 binding sites (Gasch *et al.*, 2000, Fauchon *et al.*, 2002). On average, Cbf1 binding sites were more strongly associated with gene activation than Met31 or Met32 binding sites. Strikingly, Cbf1 and Met31 or Met32 binding sites were associated with a synergistic effect on gene activation, but only when the Cbf1 binding site was found upstream.

High cooperativity between Cbf1 and Met31 or Met32 binding sites spaced at least 18 bp apart

We predicted that efficient recruitment of Met4 to the promoters of sulfur utilization genes should depend on the spacing between Cbf1 and Met31 or Met32 binding sites. To investigate the effect of varied spacing on reporter gene activation, we constructed a set of promoter libraries that differed by 2-bp increments from 6 bp to 34 bp, as well as 5-bp increments from 40 bp to 50 bp. Each promoter library had a fixed size but degenerate nucleotide sequences between the Cbf1 and Met31 or Met32 binding sites. The binding sites were flanked by 10 bp of sequence from the *MET16* promoter of *S. bayanus*, which lacks an adjacent Gcn4 site. By pooling hundreds of yeast transformants for each library, we reasoned that the contribution of nucleotide composition on Met4 recruitment and subsequent gene activation would be averaged out. We thus anticipate that growth rates for each promoter library should largely represent the aggregate effect of a certain distance on reporter gene expression.

Measurements of the growth rates of such pooled populations determined that synergistic activation required a minimum distance between Cbf1 and Met31 or Met32 binding sites (Figure 4.5). As a negative control, yeast harboring promoter libraries of varying sizes grew at similar rates in the absence of 3-AT, indicating low levels of leaky transcription from the reporter construct. Expression levels of the *HIS3* reporter gene were titrated with the addition of 1 mM 3-AT; similar results were obtained with different concentrations of 3-AT (data not shown). Binding sites whose centers were spaced fewer than 14 bp apart promoted weak reporter gene activation. At these close distances, Cbf1 and Met31 or Met32 may be sterically constrained from assembling a complex with Met4. Reporter gene activation rose sharply as the distance between binding sites was increased from 14 bp to 18 bp. The highest levels of gene activation were observed for pooled promoter libraries with binding sites spaced between 18 bp and 40 bp apart, suggesting that Cbf1 and Met31 or Met32 were optimally spaced within this distance range to interact with their respective docking sites on Met4, enabling higher levels of Met4 recruitment and subsequent gene activation. Helical phasing modulated the average growth rate by less than 25%: promoter libraries at a distance of 20 had an average growth rate of 0.36 doublings per hour, whereas promoter libraries at a distance of 26 had an average growth rate of 0.29 doublings per hour (Figure 4.5). Although promoter libraries with distances greater than 40 bp showed a gradual decrease in reporter gene activation, the average growth rates observed were still higher than for promoter libraries with distances fewer than 16 bp.

Figure 4.5) Average growth rates for pooled sequence libraries with defined distances between Cbf1 and Met31 or Met32 binding sites



Growth rates were measured for pooled transformants containing single-copy plasmids with the indicated distances between Cbf1 and Met31 or Met32 binding sites. The average growth rate and standard error of the mean are plotted for three independent trials.

Sequence context between binding sites can inhibit gene activation

In addition to characterizing the aggregate effects of binding site spacing, we also examined the effects of different spacer sequences on reporter gene activation. We assayed the growth rates for individual yeast transformants on solid media containing 10 mM or 25 mM 3-AT. Each transformant harbored a promoter with a different, random sequence between the Cbf1 and Met31 or Met32 binding sites. We observed reproducible variability in growth rates among transformants with the same distance, but different spacer sequences, between Cbf1 and Met31 or Met32 binding sites (Figure 4.6).

At each distance surveyed, a certain proportion of intervening sequences was compatible with reporter gene expression. Since the pooled growth rates in liquid media were qualitatively similar over this distance range, we interpret these proportions as the probability that a random intervening sequence would support gene activation at a given distance. At a distance of 12 bp between sites, less than 30% of the sequences supported reporter gene activation. At distances between 16 and 50 bp, the proportion of transformants that showed moderate to high levels of growth on 25 mM 3-AT varied from 38% to 60%. We observed a modest dependence of this proportion on helical phasing in the distance between binding sites.

To investigate what features of spacer sequences correlated with gene activation, we sequenced a sample of promoters with distances of 12 bp and 20 bp between the Cbf1 and Met31 or Met32 binding sites (Table 4.2). Promoters that supported gene activation (positives) were similar in nucleotide composition to promoters that inhibited gene activation (negatives). In addition, we could not find trimers or tetramers that were enriched in the positive or negative promoter sets. The most discriminating feature of

Figure 4.6) Different sequences between Cbf1 and Met31 or Met32 binding sites show a range of reporter gene activation

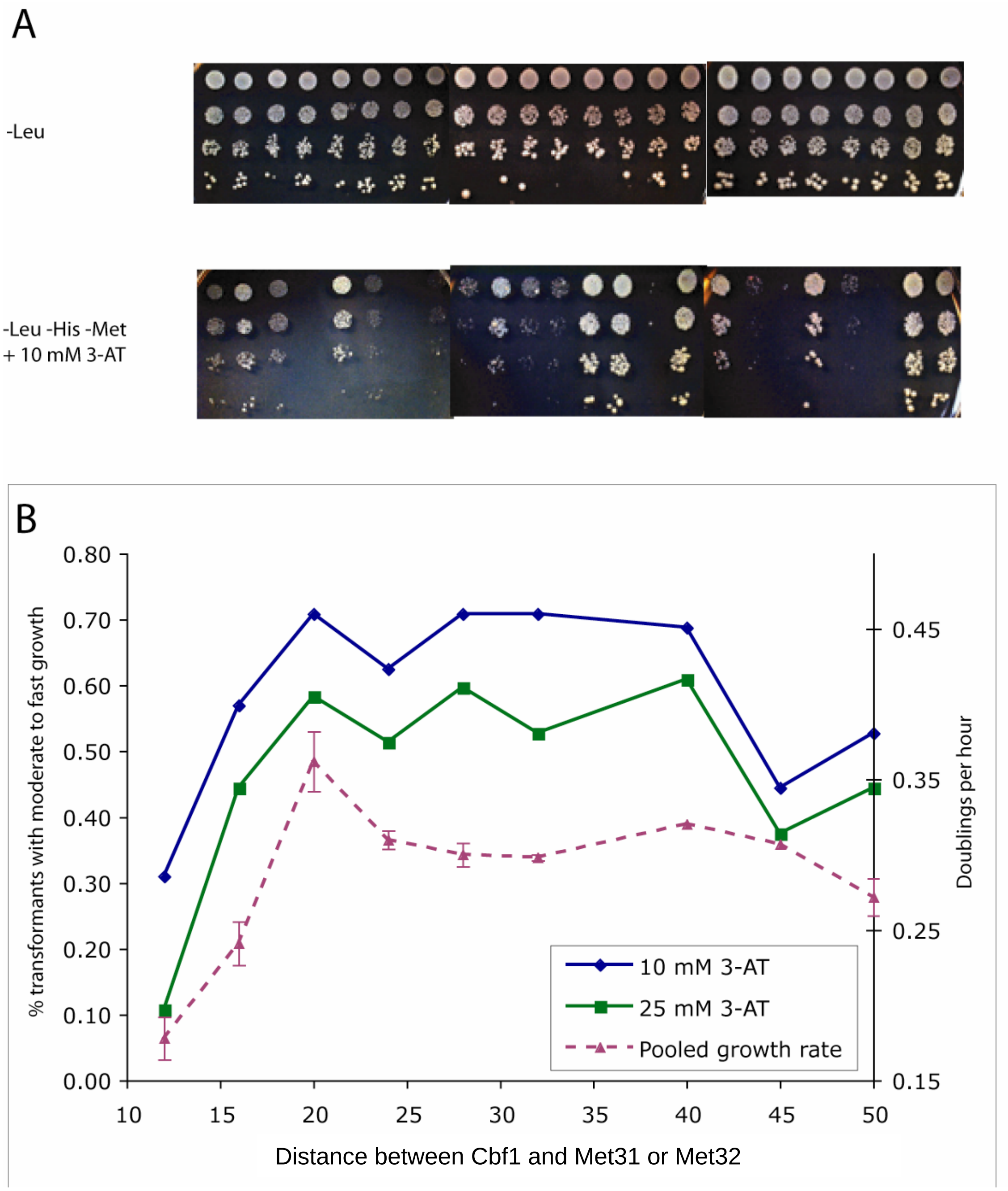


Figure 4.6 (continued)

(A) Serial dilutions of yeast containing reporter plasmids with the same distance between binding sites, but different spacer sequences. Yeast strains were grown on the indicated media for 5 days at 30°C. (B) Proportions of transformants that displayed moderate to high levels of growth on solid media with 10 mM or 25 mM 3-AT. For each distance between binding sites, growth rates of 72 transformants with different spacer sequences were assayed with serial dilutions. Average growth rates in liquid media with 1 mM 3-AT, as in Figure 3, is also shown for comparison purposes.

Table 4.2) Promoter sequences associated with reporter gene activation

(A) Center-to-center distance of 12 bp

Clone	Growth rate	Intervening sequence
5	-	TGGGTT
8	-	GAGGCG
20	-	GAGCAT
22	-	TGGATG
30	-	GTGAGT
32	-	AAAGAG
33	-	GTGACT
35	-	TGGTGT
36	-	AGAATG
44	-	TGCACT
47	-	AATTGG
48	-	AAACTC
3	+	TAAGAG
6	+	CATAGT
19	+	CGGTCC
25	+	GTTAAT
42	+	CGTCGT
2	++	CGCGTT
4	++	AACCGC
7	++	TGAGGC
27	++	CAAACG
43	++	CCATGG
46	++	GGTTGT
18	+++	ATTGGC
23	+++	AGGCAA
29	+++	ATATAT
31	+++	AAATGA
34	+++	TTGTGA

(B) Center-to-center distance of 20 bp

Clone	Growth rate	Intervening sequence
2	-	GAGTCTGATGGTCT
7	-	TGGGTTGTCAACGG
23	-	GGGCAATCGCGATG
28	-	CGTGGGGTGCTTAG
31	-	TATAAGGCGTTGGG
47	-	CGAGGGGAAAACAG

48	-	TGAGGAGATGAAGT
68	-	GAAGTGAGGAGCGG
69	-	AAGAATTACCCGGT
3	+	CCTGATGCCTACAG
16	+	CAAGGCTAGGAGCG
24	+	GCGCAGGATCGGCT
67	+	GGGTGTGAAGGGCT
9	++	TGAGCTCTTGACAT
14	++	GGTTCAACGTTACT
30	++	GCAAGGAGCGAGGG
32	++	AGGGGAACGGAGAG
33	++	TAGTGGGATTTGCG
34	++	GGTGACTAGGCCTC
46	++	GAAGTGGATTGCGT
5	+++	GGACGTAATTTCAA
8	+++	TTTACAAACTAGGG
10	+++	CGATGTACTGCCAA
11	+++	GTTTGTTGGGATGG
12	+++	GGCATTTATGGGAA
13	+++	CCCTTCCTGTGGGC
15	+++	GGTGGTTCATGGGA
18	+++	CGCGCGGGCGTCTT
25	+++	TCAGGGTTCAGCCA
26	+++	CGCGCCGAACGGGC
27	+++	TAGTGTCGGGGGCG
29	+++	GTGGTAGACGCTGC
35	+++	TTATGGTACCACCA
36	+++	TCATGCGTCGTACG
37	+++	TTGCTGGCAAGGAT
38	+++	AAAAGAGGAGATTC
39	+++	ATGTCGTCATGTGT
40	+++	AATGGCATGCTGCG
41	+++	AGAGGCAGTATCAA
43	+++	GTTTGGGTCCGGGC
72	+++	GGCATTTATGGGAA

negative promoters was a shared guanine or thymine immediately 5’ to the Met31 or Met32 binding site in 15 of 17 examples of distance 12, as well as in all 13 examples of distance 20 (Figure 4.7A). However, about half of the positive examples contained a guanine or thymine at that position, as expected.

We searched for additional residues that could discriminate among sequences that shared a guanine or thymine at the most 3’ position of the spacer region. We computed the information content at each residue within the positive and negative examples of spacer sequences using WebLogo (Schneider and Stephens, 1990; Crooks *et al.*, 2004). By focusing on the three most informative positions, we derived nucleotide combinations that predicted negative promoters with an overall sensitivity of 80% and a specificity of 89% (Table 4.3).

To test whether the A_{11}-T_{17} nucleotide combination was sufficient to inhibit gene activation, we identified five promoter sequences with a B_{11}-T_{17} combination and converted the nucleotide at position 11 to an adenine by site-directed mutagenesis. For four of the five strains, similar levels of reporter gene activation were driven by the original and mutant promoters, as assayed by serial dilutions on media containing 10 mM or 25 mM 3-AT (Figure 4.7B). Only strain #14 showed decreased reporter gene activation by the mutated promoter in the presence of 10 mM 3-AT. Thus, the effects of sequence context are not encoded by individual positions within the primary nucleotide structure.

Figure 4.7) Discriminating nucleotides fail to inhibit reporter gene activation *in vivo*

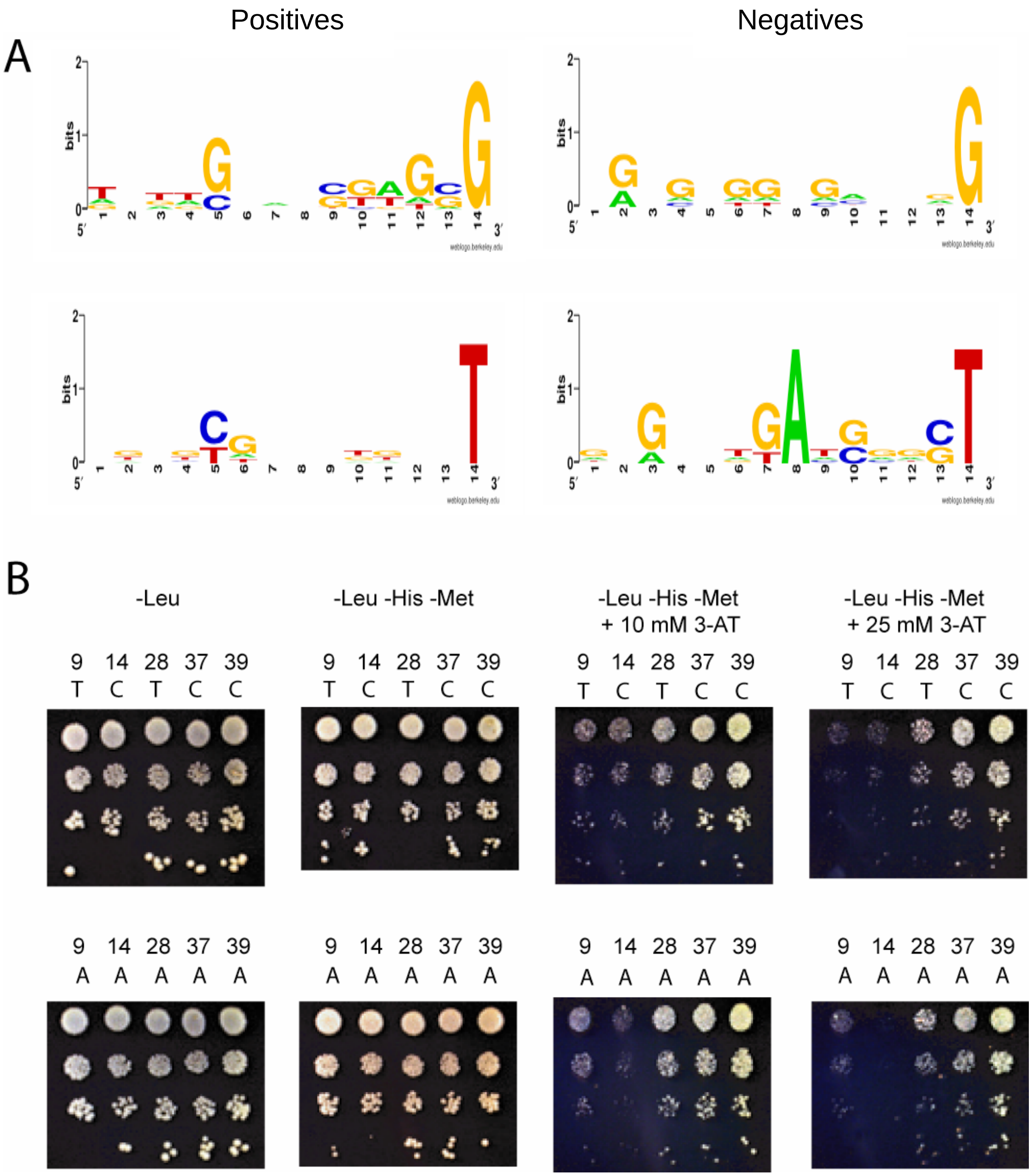


Figure 4.7 (continued)

(A) Sequence logos of intervening sequences between Cbf1 and Met31 or Met32 binding sites were generated with WebLogo (Crooks *et al.*, 2004). At each position, the height of the nucleotide corresponds to its frequency in the sequenced sample. (B) Serial dilutions of yeast containing reporter plasmids with or without an adenine at position 11.

Table 4.3) Nucleotide combinations that correlate with lack of reporter gene activation

Size	Sequence combination	Sensitivity	Specificity
12	R_5-G_9	6 / 6 (100%)	6 / 7 (86%)
12	K_6-T_9	8 / 9 (89%)	8 / 9 (89%)
20	W_8-G_{17}	5 / 8 (62%)	5 / 5 (100%)
20	A_{11}-T_{17}	5 / 5 (100%)	5 / 6 (83%)

IUPAC symbols: R = A or G; K = A or C; W = A or T.

DISCUSSION

Promoter architecture features of yeast sulfur utilization genes

We have developed a synthetic promoter assay to test how various features of promoter architecture affected activation of a *HIS3* reporter gene in the context of a common minimal promoter. We applied this system to characterize the collaborative recruitment of the coactivator, Met4, by the transcription factors, Cbf1 and Met31 or Met32, in response to methionine starvation. The relative order of binding sites was important, since a Cbf1 binding site was required upstream of a Met31 or Met32 binding site for full activation. Two Cbf1 binding sites could moderately activate reporter gene expression, yet the mechanism for this enhanced activation is unclear. Synergistic activation of reporter gene expression occurred when Cbf1 and Met31 or Met32 binding sites were spaced at least 18 bp apart. Notably, the minimum distance required for coactivator recruitment is further than the maximum range of cooperatively binding transcription factors. Finally, we discovered that different sequence contexts between binding sites produced considerable heterogeneity of reporter gene activation.

These promoter architecture requirements reveal insights on regulatory mechanisms for the coactivator, Met4. The influence of Cbf1 and Met31 or Met32 binding site order on reporter gene activation implies that the spatial orientation of the Met4 activation domain is required for the recruitment of downstream targets. However, the rather flexible distance constraints between binding sites suggests that Met4 recruitment may not require simultaneous protein-protein interactions among the bound transcription factors. Intriguingly, the recruitment of Met4 to a common minimal promoter seems to depend more on the sequence context between Cbf1 and Met31 or

Met32 binding sites than on the distance between them. In light of these results, previous studies that varied distances between transcription factor binding sites should be reassessed, since they usually considered only a single sequence context for each distance. In addition, promoter classification programs should strive to incorporate sequence context effects into their algorithms.

Possible effects of sequence context between transcription factor binding sites

Sequence context could alter Met4 recruitment in several ways. First, residues adjacent to binding sites could reduce the binding affinity of Cbf1 or Met31 or Met32. Accordingly, we found that all spacer sequences that inhibited reporter gene activation contained a guanine or thymine immediately 5' to the Met31 or Met32 binding site. Secondly, the DNA bendability of the spacer sequence could alter the conformation of Cbf1, which bends DNA by approximately 68° (Niedenthal *et al.*, 1993). Conformational changes in Cbf1 could affect its protein-protein interactions with Met28 or Met4, thus reducing Met4 recruitment. A requirement for DNA bendability on protein-protein interactions has been recently shown for the transcription factor, Mcm1, which bends DNA by 66°, comparable to the bend angle induced by Cbf1 (Lim *et al.*, 2003). A point mutant in Mcm1 with a DNA bending angle of 46° had a lower affinity for cooperative binding with Fkh2 than a mutant with a DNA bending angle of 49°, suggesting that a certain threshold of DNA bending was required for ternary complex formation *in vitro* (Lim *et al.*, 2003). Circular permutation assays on promoters with different sequence contexts could test whether the extent of bendability correlates with reporter gene activation. In addition, chromatin immunoprecipitation studies could identify the transcription factors whose binding *in vivo* is affected by sequence context.

Whereas the influence of sequence context on gene activation has been widely reported (*e.g.*, Elledge and Davis, 1989; Mai *et al.*, 2000), the key determinants of sequence context have been poorly defined. Except for the residue adjacent to the Met31 or Met32 binding site, we could not find features of the primary nucleotide structure that correlated with gene activation. Previous studies have reported that protein-DNA interactions can be affected by physicochemical properties of DNA, such as twist (Olson 1995). Although we assessed several dinucleotide parameters, we could not find any significant correlation between the average parameter value of a spacer sequence and reporter gene activation (data not shown) (Olson *et al.*, 1998).

A couple of follow-up experiments could better characterize key features of sequence context. Whereas it is not feasible to test all possible sequence variants between a pair of transcription factor binding sites, a sample of several hundred different sequences would provide more statistical power to infer key determinants of sequence context. Fluorescence activated cell sorting of yeast cells that expressed green fluorescent protein would provide a higher throughput method to assay the effect of sequence variants on reporter gene expression. Another approach could investigate the minimum number of nucleotide changes that render a promoter unable to drive reporter gene expression. We have identified several pairs of active and inactive sequences that differ by six nucleotides. Therefore, growth assays for all 64 possible recombinants between the sequence pairs would indicate which positions are necessary for reporter gene activation.

In order to sample a large number of promoter architectures, we assayed reporter gene expression from a single-copy plasmid, which yields over 10,000-fold higher transformation efficiency than chromosomal integration. We have not explored how the flanking sequence composition of wild-type promoters may affect the basal or Met4-induced nucleosomal accessibility of Cbf1 and Met31 or Met32 binding sites in the genome. Cbf1 can also modulate nucleosome positioning and recruit the Isw1 chromatin remodeling complex (Moreau *et al.*, 2003; Kent *et al.*, 2004). Thus, additional determinants of local sequence context that affect the binding or DNA bending of Cbf1 may influence Met4 recruitment and gene activation in a chromosomal context. By integrating minimal promoters into the *MET16* locus of an *isw1Δ* mutant strain, we could dissect the relative contribution of chromatin remodeling on Met4 recruitment.

CHAPTER 5

REGULATORY DIVERSIFICATION OF

TRANSCRIPTION FACTOR PARALOGS

PREFACE

Transcriptional regulatory networks have two types of components: transcription factors and target genes. In the previous chapter, I described experiments to characterize how promoter architecture influences the regulation of yeast sulfur utilization genes. In this chapter, I focus on the transcriptional and post-translational regulation of the duplicated transcription factor pair, Met31 and Met32. I also developed a computational approach to systematically evaluate theories about the evolution of transcription factor paralogs.

BACKGROUND

Gene expression studies of different yeast species have revealed that different groups of genes are co-expressed in response to similar environmental conditions (Tsong *et al.*, 2003, Rustici *et al.*, 2004). However, the mechanisms by which groups of genes become regulated by different transcription factors are poorly characterized. One possible mechanism invokes the gain or loss of binding sites in the transcriptional regulatory regions of different genes that are regulated by the same transcription factor (Wray *et al.*, 2003). Moreover, amino acid changes in transcription factors may alter their regulatory specificity in various ways (Hsia and McGinnis, 2003). Changes to the DNA binding domain may alter a transcription factor's sequence specificity and thus its affinity for binding sites in different target genes (Gasch *et al.*, 2004). In addition, the gain of protein-protein interactions among transcription factors may recognize composite elements and impose a novel regulatory combination in transcription control regions that can be subject to selection (Lohr *et al.*, 2001).

Duplication of transcription factors creates regulatory redundancy, which may relax selection pressures on amino acid sequences. In the adaptive evolution model, one of the transcription factor paralogs retains the ancestral regulatory function, whereas the other paralog may either be lost by mutation or gain a new function at a low frequency (Ohno, 1970). However, a surprisingly high fraction of transcription factor paralogs are retained in extant species without obviously new functions, despite adequate evolutionary time for the loss of a paralog to occur (reviewed in Force *et al.*, 1999). This conundrum led to the proposal of the subfunctionalization model, which speculates that each paralog retains subsets of the ancestral functions (Force *et al.*, 1999). Both paralogs would thus be required to fully complement the ancestral transcription factor's regulatory functions. Notably, some transcription factor families have been amplified in certain clades, such as: the zinc binuclear cluster in fungi (Akache *et al.*, 2001); nuclear hormone receptors in metazoans (Escriva *et al.*, 1997); MADS box family in plants (Becker and Theissen, 2003). It is tempting to speculate that the amplification of transcription factor families may enable the evolution of specialized regulatory networks that lead to phenotypic diversity.

Experimental studies on several pairs of duplicated transcription factors in *Saccharomyces cerevisiae* concur that these paralogs have overlapping, but distinct, functions. Although deletions of single transcription factors are often viable due to partial complementation by the other paralog, genetic and biochemical studies often indicate that each paralog can have different roles. For instance, the winged helix transcription factors, Fkh1 and Fkh2, share 47% sequence identity and identical DNA binding specificities (Hollenhorst *et al.,* 2000; Zhu *et al.*, 2000). Fkh1 regulates donor type switching by binding to the recombination enhancer (Sun *et al.*, 2002), whereas Fkh2 interacts with Mcm1p to regulate cell cycle genes (Boros *et al.*, 2003). These proteins also have opposing roles in regulating transcriptional elongation and termination (Morillon *et al.*, 2003). In another example, the zinc finger transcription factors, Met31 and Met32, are 46% identical at the amino acid level, though they bind the same recognition sequence with different affinities (Blaiseau *et al.*, 1997). Genetic studies suggest different roles for the paralogs with respect to the ubiquitin ligase, Met30, which is a negative regulator of Met4. A *met30Δ met31Δ* strain is synthetically lethal, whereas a *met30Δ met32Δ* mutant is a methionine prototroph (Patton *et al.*, 2000). Taken together, these studies suggest that transcription factor paralogs may attain divergent functions by interacting with different proteins.

I sought evidence for distinct regulatory functions between duplicated transcription factor pairs in *Saccharomyces cerevisiae*. For each pair of duplicated transcription factors, I counted the number of amino acid substitutions that occurred in the divergence of each paralog from an outgroup single-copy homolog. I looked for cases in which one paralog had accumulated statistically more substitutions that the other paralog. In addition, I investigated whether Met31 and Met32 interacted with different proteins *in vivo*.

MATERIALS AND METHODS

Sequence alignments for duplicated transcription factors

The published genome sequences of *Kluyveromyces waltii* (Kellis *et al.*, 2004) and *Ashbya gossypii* (Dietrich *et al.*, 2004) were obtained from Genbank. Both genome sequencing projects assigned homology between each predicted open reading frame and a corresponding gene from *Saccharomyces cerevisiae*. Thirty-one pairs of duplicated transcription factors in *S. cerevisiae* were matched with orthologs in *K. waltii* and *A. gossypii*. Amino acid sequences of paralogous pairs were aligned with T-COFFEE, along with the homolog from either *K. waltii* or *A. gossypii* included as an outgroup sequence (Notredame *et al.*, 2000). The coordinates of DNA binding domains in each amino acid sequence were annotated by the Pfam web server (Bateman *et al.*, 2004).

Tajima relative rates test for accelerated evolution

The Tajima relative rates test evaluates whether differential rates of evolution has occurred along one species branch (Tajima, 1993). This test is essentially a chi-square test with the null hypothesis that the number of amino acid changes along each branch are equivalent. From the amino acid alignment, let m_1 represent the number of amino acid differences between an outgroup sequence and one of the paralogs; let m_2 represent the number of amino acid differences between the outgroup sequence and the other paralog. Gapped positions were excluded from this analysis. The Tajima test statistic is given by:

$$\chi^2 = \frac{(m_1 - m_2)^2}{m_1 + m_2}$$

which is chi-square distributed with one degree of freedom. Correction for multiple testing involved controlling the False Discovery Rate with $q < 0.01$ (Benjamini and Hochberg, 1995). Significantly large values of the test statistic would reject the null hypothesis, and thus indicate that more mutations have accumulated along one branch than expected by chance.

Maximum-likelihood estimates of sequence divergence (PAML)

Another test for differential rates of evolution reconstructs the sequence of the common ancestor by maximum likelihood under a model for sequence evolution. Given the three-way amino acid alignments, the PAML software package infers the rate of amino acid divergence along each branch from the common ancestor to each extant sequence (Yang, 1997). Significantly large rate differences indicate changes in the rate of evolution along the corresponding branch.

Yeast cell extracts from Met31^{FLAG} and Met32^{FLAG} overexpression strains

Yeast strain YMT235 was transformed with plasmids containing *MET31* or *MET32* genes with C-terminal FLAG epitope tags under the control of a galactose-inducible promoter (Ho *et al.*, 2002). Cells were grown to early log phase (OD_{600} = 0.25) in 500 mL of sulfur-free B media with 50 μM methionine and 2% raffinose (Cherest and Surdin-Kerjan, 1992). At this point, epitope-tagged transcription factors were induced by adding galactose to a final concentration of 2%. After 1 hour, the cultures were split in half, and methionine was added to one culture to a final concentration of 1 mM. Cells were pelleted and frozen in liquid nitrogen after resuspension in 1 mL low-salt RadioImmunoPrecipitationAssay (RIPA) buffer (50 mM Tris pH 7.5, 50 mM NaCl, 1% deoxycholic acid, 1% Triton X-100, 0.1% SDS, 10 mM sodium pyrophosphate, 5 mM EDTA, 5 mM EGTA, 50 mM sodium fluoride, 0.1 mM orthovanadate, 1 mM PMSF, 2 μg/mL protease inhibitors and 50 μg/mL ethidium bromide). Whole cell extracts were recovered after cell lysis by vortexing with glass beads. Total protein recovery was assayed by Bradford reagent.

Co-immunoprecipitation

For immunoprecipitation of Met31^{FLAG} or Met32^{FLAG}, 5 mg of total protein was incubated with 50 μL of anti-FLAG M2-agarose affinity beads (Sigma) at 4°C for 1.5 hours. After 3 washes with 1 mL low-salt RIPA buffer, bound proteins were resuspended in 10 μL of sample buffer and denatured by boiling for 5 min. Protein samples were resolved on 10% polyacrylamide gels and transferred to PVDF membranes. Western blots were conducted with rabbit polyclonal antibodies to Met28 or Met4 (a gift from Traci Lee and Mike Tyers). The anti-Met28 antibody was diluted to 1:200, and the anti-Met4 antibody was diluted 1:1000 in Tris-buffered saline (10 mM Tris pH 7.5, 150 mM NaCl) with 0.05% Tween-20 and 5% milk. Antibodies were incubated for 1.5 hr at room temperature, followed by 3 washes with Tris-buffered saline with 0.05% Tween-20. A 1:10,000 dilution of donkey anti-rabbit antibody conjugated with horseradish peroxidase was incubated for 1.5 hr at room temperature, followed by 3 washes in Tris-buffered saline with 0.05% Tween-20, then 2 washes in Tris-buffered saline without Tween-20. Antibody staining was conducted with SuperSignal West Pico Chemiluminescent Substrate (Pierce).

RESULTS

Some transcription factor paralogs show differential rates of evolution

Each pair of duplicated transcription factors in *S. cerevisiae* was aligned with their single-copy homolog from either *K. waltii* or *A. gossypii* as an outgroup sequence. Two computational tests evaluated whether differential rates of sequence evolution occurred along a branch leading to one of the *S. cerevisiae* paralogs. Among 31 transcription factor paralogs tested, the Tajima test identified 13 pairs with significantly different rates of evolution when the *K. waltii* ortholog was used as the outgroup sequence, of which 8 pairs also showed significantly different rates of evolution when the *A. gossypii* ortholog was used as the outgroup sequence (Table 5.1). To assess whether these changes may affect the binding specificity of the transcription factor paralog, this analysis was repeated on alignments of the DNA binding domains only. Only 4 pairs showed significantly different rates of evolution when the *K. waltii* ortholog was used as the outgroup sequence, of which 3 were also significant using *A. gossypii* as the outgroup (Table 5.2). Since DNA binding domains are typically short (less than 60 amino acids) and the most highly conserved regions of transcription factors, the lower number of significant differences may simply reflect decreased power of the chi-square test. Finally, results from the maximum likelihood inference of branch lengths with PAML largely agreed with the Tajima test for accelerated evolution (Table 5.1).

Table 5.1) Tajima test for differential rates of evolution on full-length proteins

				# of amino acid changes from *K. waltii* sequence			PAML branch length from common ancestor; *K. waltii* outgroup	
Gene 1	Gene 2	Tajima χ^2	Ungapped length	Gene 1	Gene 2	*p*-value	Gene 1	Gene 2
HMS2	*SKN7*	74.3	264	**110**	14	**7×10^{-18}**	**1.61**	0.28
EDS1	*RGT1*	73.5	756	**180**	50	**1×10^{-17}**	**1.15**	0.33
GIS1	*RPH1*	30.5	602	**145**	65	**4×10^{-8}**	**0.68**	0.35
CAD1	*YAP1*	22.7	295	80	30	2×10^{-6}	1.63	0.67
MIG1	***MIG2***	20.3	254	21	**62**	**7×10^{-6}**	0.72	**1.96**
RLM1	***SMP1***	18.5	372	17	**53**	**2×10^{-5}**		
ECM22	*UPC2*	15.1	664	**93**	47	**1×10^{-4}**	**0.50**	0.29
CUP2	*HAA1*	12.5	136	35	11	4×10^{-4}	1.31	0.42
MSN2	*MSN4*	11.9	439	27	59	6×10^{-4}	0.49	0.81
ACE2	***SWI5***	11.2	566	58	**100**	**8×10^{-4}**	0.50	**0.85**
DIG1	*DIG2*	10.6	236	21	48	1×10^{-3}	0.80	1.23
FKH1	*FKH2*	8.1	226	**41**	19	**5×10^{-3}**	**0.54**	0.28
AFT2	*RCS1*	7.9	297	52	27	5×10^{-3}	0.80	0.51
YML081W	*ZMS1*	6.9	1040	106	148	9×10^{-3}	0.41	0.54
NRG1	*NRG2*	5.6	113	4	14	0.02	0.11	0.57
CIN5	*YAP6*	4.5	163	8	19	0.03	0.46	0.78
YHP1	*YOX1*	4.1	260	35	20	0.04	0.65	0.46
ACA1	*CST6*	3.4	217	27	15	0.07	0.48	0.24
YBP1	*YBP2*	3.3	566	91	68	0.07	0.81	0.62
MET31	*MET32*	2.5	125	17	9	0.12	0.66	0.38
DAL80	*GZF3*	1.4	203	14	21	0.24	0.38	0.67
NHP6A	*NHP6B*	1.0	33	0	1	0.32	0.06	0.04
STP3	*STP4*	0.3	208	15	12	0.56	0.28	0.23
SUT1	*SUT2*	0.0	204	24	25	0.89	0.56	0.60
YKL222C	*YOR172W*	0.0	638	93	94	0.95		

Bold entries denote the transcription factor paralog with differential rates of evolution as assessed with the Tajima χ^2 statistic ($q < 0.01$ after correction for multiple testing), compared to both *K. waltii* and *A. gossypii* as outgroup sequences.

Table 5.2) Tajima test for differential rates of evolution on DNA binding domains

				# of amino acid changes from *K. waltii* sequence			PAML branch length from common ancestor: *K. waltii* outgroup	
Gene 1	Gene 2	Tajima χ^2	Ungapped length	Gene 1	Gene 2	*p*-value	Gene 1	Gene 2
HMS2	*SKN7*	27.6	167	**55**	12	**2×10^{-7}**	**1.24**	0.41
CUP2	*HAA1*	11	40	**11**	0	**1×10^{-3}**	**0.34**	0
FKH1	*FKH2*	8.9	100	**16**	3	**3×10^{-3}**	**0.30**	0.07
MIG1	***MIG2***	7.4	53	1	**10**	**7×10^{-3}**		
EDS1	*RGT1*	3.6	41	6	1	0.06	0.66	0.14
MSN2	*MSN4*	2.7	52	5	1	0.12	0.24	0.11
ACE2	*SWI5*	2.7	83	5	1	0.12	0.14	0.09
NRG1	*NRG2*	2	53	0	2	0.16	0.02	0.22
DAL80	*GZF3*	1	35	3	1	0.32	0.18	0.08
MET31	*MET32*	1	53	3	1	0.32	0.23	0.11
STP3	*STP4*	1	23	0	1	0.32	0	0.09
YHP1	*YOX1*	1	57	6	3	0.32	0.22	0.16
ACA1	*CST6*	0.8	65	7	4	0.37	0.27	0.17
OAF1	*PIP2*	0.7	39	2	4	0.41	0.11	0.30
ECM22	*UPC2*	0.2	39	3	2	0.66	0.15	0.15
GIS1	*RPH1*	0	26	1	1	1	0.04	0.04
NHP6A	*NHP6B*	0	69	3	3	1	0.06	0.05
PHD1	*SOK2*	0	81	0	0	1		

Bold entries denote differential rates of evolution ($q < 0.01$ after correction for multiple testing) between the transcription factor paralogs, compared to both *K. waltii* and *A. gossypii* as outgroup sequences.

Literature analysis of transcription factor paralogs with differential rates of evolution

Several transcription factor pairs demonstrated diffrential rates of evolution of one paralog, using both *K. waltii* and *A. gossypii* as outgroups. The *HMS2* and *SKN7* pair of transcription factors showed the most significant difference in evolutionary rates: over 100 amino acids changes occurred between *HMS2* and the *K. waltii* homolog, whereas only 14 changes occurred between *SKN7* and the outgroup sequence. Strikingly, these changes correspond to the loss of a response regulator domain in *HMS2* that is present in both *SKN7* and the homologs from *K. waltii* and *A. gossypii* (Bateman *et al.*, 2004). The functional inactivation of an entire structural domain lends anecdotal support to the subfunctionalization theory, which predicts that certain functions of the ancestral sequence may be lost in one of the paralogs.

Three pairs belong to the classical C_2H_2 zinc finger family: *GIS1-RPH1*, *MIG1-MIG2* and *ACE2-SWI5*. All of these paralogs have significantly different amino acid changes outside of their DNA binding domains. Notably, previous studies have demonstrated that Ace2 and Swi5 have different interaction partners. The homeodomain transcription factor, Pho2, can directly interact with Swi5, but not Ace2 (Dohrmann *et al.*, 1996). Conversely, the cyclin-dependent kinase, Cdc28, can phosphorylate Ace2, but not Pho5 (O'Conallain *et al.*, 1999).

Two pairs of transcription factor paralogs belong to the fungal-specific zinc binuclear cluster family: *EDS1-RGT1* and *ECM22-UPC2*. Recent studies have shown different interactions between Hap1 with Ecm22 and Upc2 at ergosterol-regulated promoters (Brandon Davies & Jasper Rine, UC Berkeley, personal communication).

Two other pairs are the only representatives of their respective transcription factor families. Experimental characterization of Fkh1 and Fkh2 has been discussed in the Introduction. Given the characterized roles of these transcription factors, it is possible that Fkh2 retains the ancestral function of cell-cycle regulation, whereas the accelerated evolution of Fkh1 has enabled it to acquire a role at the recombination enhancer. Finally, Rlm1 and Smp1 are MADS box transcription factors involved in response to cell wall stress that have not been well-characterized (Dodou and Treisman, 1997).

Met31 and Met32 interact with different proteins *in vivo*

In addition to a genome-wide computational survey for different evolutionary rates between transcription factor paralogs, I was also interested in characterizing different functional roles for the regulators of sulfur utilization genes, Met31 and Met32. To assess whether Met31 and Met32 have different interaction partners *in vivo*, epitope-tagged versions of these transcription factors were overexpressed in both methionine-starved and methionine-replete conditions. Met31^{FLAG}- and Met32^{FLAG}-containing complexes were recovered from whole cell extracts by immunoprecipitation with anti-FLAG agarose beads. Immunoblots with antibodies to Met4 demonstrated that both Met31 and Met32 interacted with Met4 *in vivo* (Figure 5.1A). The multiple bands correspond to Met4 isoforms conjugated to various numbers of ubiquitin molecules (Kaiser *et al.*, 2000). In contrast, antibodies to Met28 only detected an interaction with Met32 (Figure 5.1B). Since Met28 stabilizes the binding of Cbf1 to DNA (Kuras *et al.*, 1997), the binding of Met32 may increase the levels of Met4 recruitment mediated by Cbf1.

Figure 5.1) Met31 and Met32 interact differentially with Met28 *in vivo*

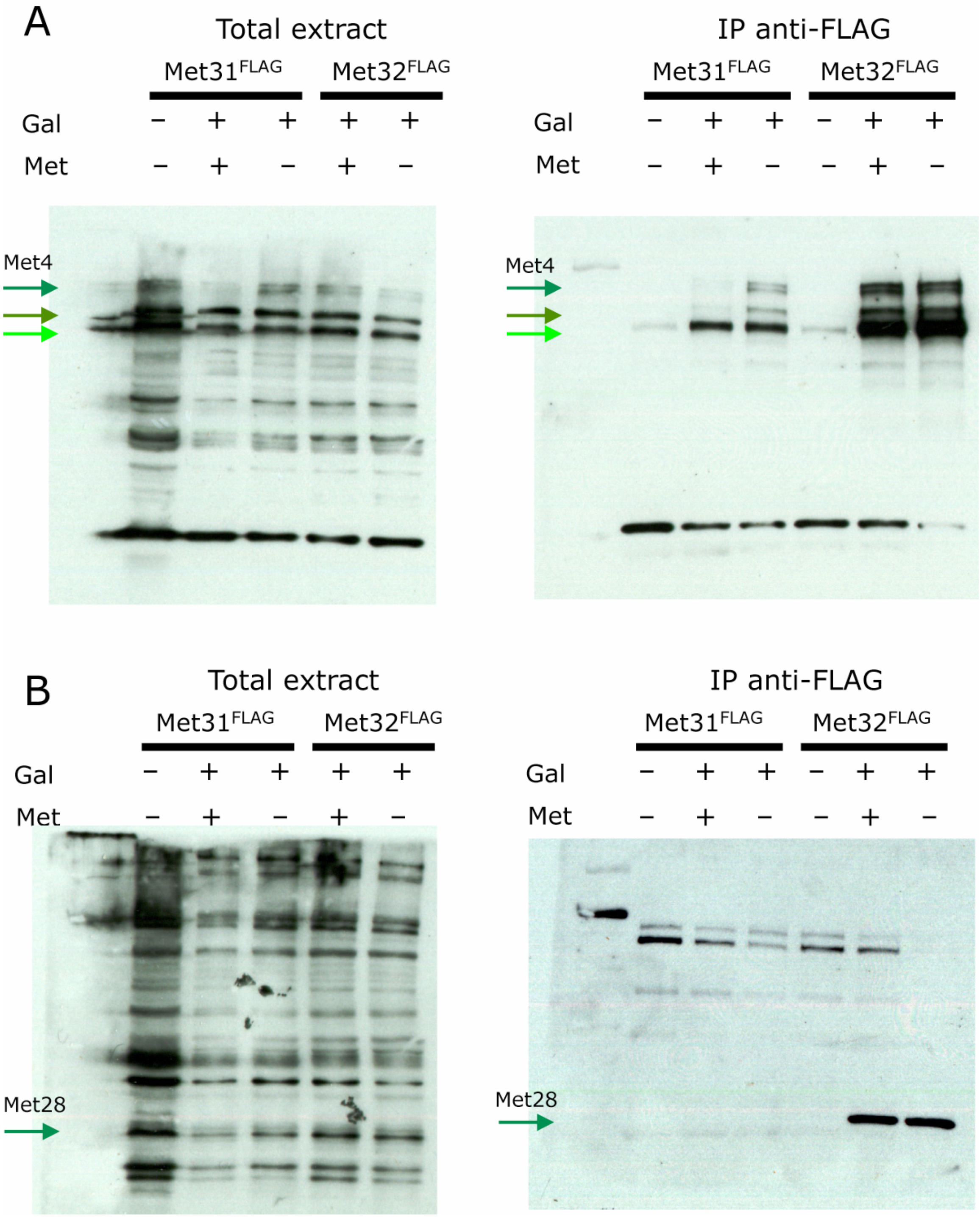


Figure 5.1 (continued)

(A) Met4 interacts with Met31^{FLAG} and Met32^{FLAG}. Western blots with rabbit polyclonal anti-Met4 antibodies were used to probe proteins from whole cell extracts and from Met31^{FLAG} or Met32^{FLAG} immunoprecipitates. Overexpression of epitope-tagged Met31^{FLAG} or Met32^{FLAG} was induced with galactose. The sulfur regulatory response was activated by limiting methionine in the growth media (– corresponds to 0.05 mM methionine), or repressed by adding methionine to 1 mM (+). Arrows denote ubiquitinated isoforms of Met4.

(B) Met28 interacts with Met32^{FLAG}, but not with Met31^{FLAG}. Same as above, except rabbit polyclonal anti-Met28 antibodies were used. Arrows denote Met28.

The transcription of *MET31* and *MET32* are differentially regulated

Promoter sequence analysis suggests that *MET31* may be constitutively expressed, whereas *MET32* is induced on sulfur limiting conditions. Binding sites for both Cbf1 and Met31 or Met32 are absent from the *MET31* promoter, but can be found in the *MET32* promoter. This hypothesis was verified by retrospective analysis of microarray data. Relative gene expression for *CBF1*, *GCN4*, *MET4*, *MET30*, *MET31* and *MET32* were compiled from previously published studies on amino acid starvation (Gasch *et al.*, 2000), 3-aminotriazole treatment (Natarajan *et al.*, 2001), *gcn4* deletion mutants (Hughes *et al.*, 2000b) and cadmium treatment (Fauchon *et al.*, 2002). Only cadmium treatment represents a specific sulfur-limiting condition, whereas the first three conditions reflect general control in response to amino acid starvation. Hierarchical clustering on this small set of genes revealed that expression levels of *MET4*, *MET28*, *MET30*, and *MET32* were induced under cadmium treatment in a Met4-dependent manner (Figure 5.2A). In contrast, *CBF1* and *MET31* showed less than two-fold expression change in multiple conditions.

Intriguingly, the strong induction of *MET28* under amino acid starvation conditions, as well as its lower expression levels in a *gcn4*Δ mutant, suggests that Gcn4 exerts feed-forward regulation on the sulfur transcriptional network (Figure 5.2B). In particular, Gcn4 activation by general amino acid control amplifies *MET28* transcription. Higher levels of Met28 should stabilize Cbf1 and Met32 binding to sulfur-regulated promoters, thus potentially amplifying gene expression.

Figure 5.2) Transcriptional feedback of sulfur regulatory network

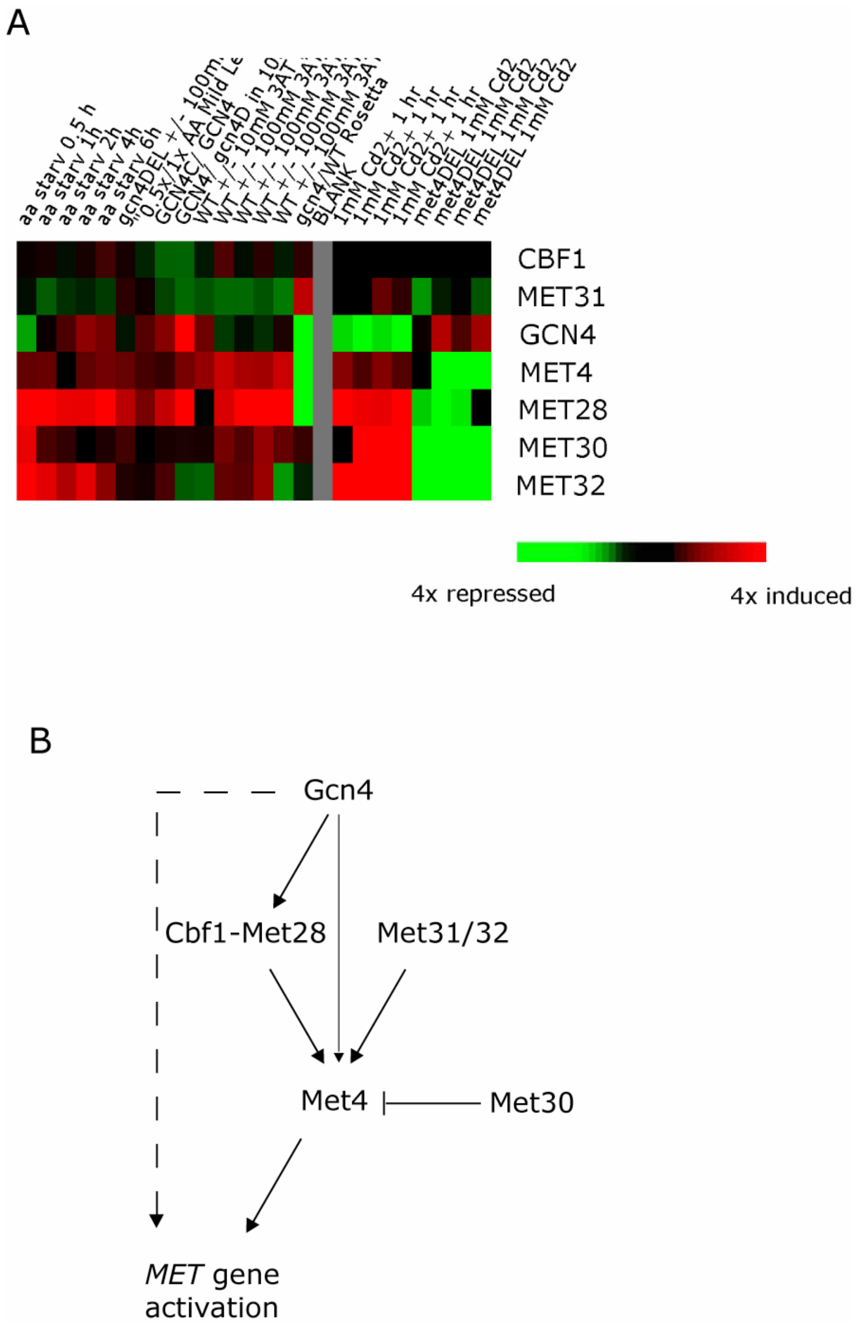


Figure 5.2 (continued)

(A) Gene expression data for transcription factors that regulate sulfur utilization genes. Each row corresponds to the labeled transcription factor, and each column corresponds to a different microarray condition. The first five columns represent an amino acid starvation timecourse (Gasch *et al.*, 2000). Columns 6 to 14 were obtained from Natarajan *et al*., 2001. Column 6 corresponds to a decrease of amino acid concentrations in half, in addition to leucine starvation; column 7 features treatment of a *gcn4*Δ strain with 100 mM 3-aminotriazole; column 8 compares a constitutive GCN4 allele with wild-type; column 9 compares a wild-type strain with a *gcn4*Δ strain; column 10 represents treatment of a wild-type strain with 10 mM 3-aminotriazole; and columns 11 to 14 are replicate treatments of a wild-type strain with 100 mM 3-aminotriazole. The last 8 columns correspond to replicate treatments of wild-type or *met4*Δ strains with 1 mM cadmium. Red pixels indicate induced genes, green pixels indicate repressed genes and pixel intensity reflects the magnitude of gene expression change. (B) Feed-forward transcription factor network. In addition to transcriptionally activating both Met4 and Met28, Gcn4 can directly bind to the promoters of some sulfur utilization genes.

Met31 and Met32 may be differentially phosphorylated

Post-translational modifications represent another possible difference between the regulation of Met31 and Met32. Indeed, a large-scale mass spectrometry effort purified phosphorylated isoforms of Met31 (Ho *et al.*, 2002; Kevin Breitkreutz and Mike Tyers, unpublished observations). Phosphorylated residues were mapped from peptide fragmentation spectra to Ser-26 or Ser-27; Ser-45 or Ser-46; and Ser155 or Ser-158 (Kevin Breitkreutz and Mike Tyers, unpublished observations). The amino acid sequence context of the Ser-46 phosphorylation site suggests that Met31 may be a substrate for the Cdc28 cyclin-dependent kinase (Yaffe et al, 2001). Notably, a multiple sequence alignment of Met31 and Met32 orthologs from five closely-related yeast species revealed that none of the putatively phosphorylated serine residues in Met31 are conserved in corresponding positions of Met32 (Figure 5.3).

Figure 5.3) Multiple sequence alignment of phosphorylated residues on Met31

```
MET31_Scer      MNVDEIFLKQAAEAIAVISSSPTHTDPIIRELLHRIRQSSP
MET31_Spar      MNVDEVFLKQAAEAIAVTSSSPTHTDPIIRELLHRIRQSSP
MET31_Smik      MNVDEIFLQQAAEAIAVTSASPTHTDPIIQELLQRIRQSSP
MET31_Skud      MNADEVFLKQAAEAIAVTSSTPSNTDPIIHELLQRIRQSSP
MET31_Sbay      MNADEVFLKQAAEAIVVTSSSTTSDPIIQELLQRIRQSSP
MET32_Scer      EDQDAAFIKQATEAIVDVSLNIDNIDPIIKELLERVRNRQN
MET32_Spar      EDQDAAFIKQATEAIVDVSLNVDNIDPIIKELLERVRNRQN
MET32_Smik      EDEDAAFIKQATEAIVDVSLNMDNIDPIIKELLERVRKRRN
MET32_Skud      EDQDSAFIKQATEAIVDVSLNVDSIDPIIKELLQRVRNMQN
MET32_Sbay      EDQDSAFIKQATEAIVDISLDINNIDPIIKELLQRVKNTRN
                   : *  *:::*.  *      :*.*:::
                   |         |         |         |
                   10        20        30        40
```

```
MET31_Scer      RHSNTLTCQRNRKKLSEGSDVDVDELIKDAIKN
MET31_Spar      RHSNTLTCQRNRKKLSEGSDVDVDELIKDAIKN
MET31_Smik      RHSNTLTCQRNRKKLSEGSDVDVDELIKDAIKN
MET31_Skud      RHSNTLTCRRNRKKLCEGSDVDVDELIKDAIKN
MET31_Sbay      RHSNTLTCQRNRKKLCEGSDVDVDELIKDAIKN
MET32_Scer      RHYDTLTCRRNRTKLLTAGGEGINELLKKVKQS
MET32_Spar      RHYDTLTCRRNRTKLLTAGGEGINELLKKVKQS
MET32_Smik      RHYDTLTCRRNRTKLLTAGGEGINELLRKVKQS
MET32_Skud      RHYDTLTCRRNRTKLLTAGGESINELLKKVKQS
MET32_Sbay      RHYDTLTCRRNRSKLLSAGGEGINELLKKVKQS
                 ::*.  ... .::::.. :.
                |         |         |         |
               140       150       160       170
```

Amino acid sequences, obtained from the MIT and Washington University genome sequencing centers, were aligned using T-COFFEE (Notredame *et al.*, 2000). In the above excerpts from the multiple sequence alignment, phosphorylated serine residues mapped by mass spectrometry on the *S. cerevisiae* protein are shown in bold. Serines conserved in orthologous sequences are also indicated in bold. Note that none of the columns aligned to phosphorylated serines are conserved in any of the *MET32* orthologs.

DISCUSSION

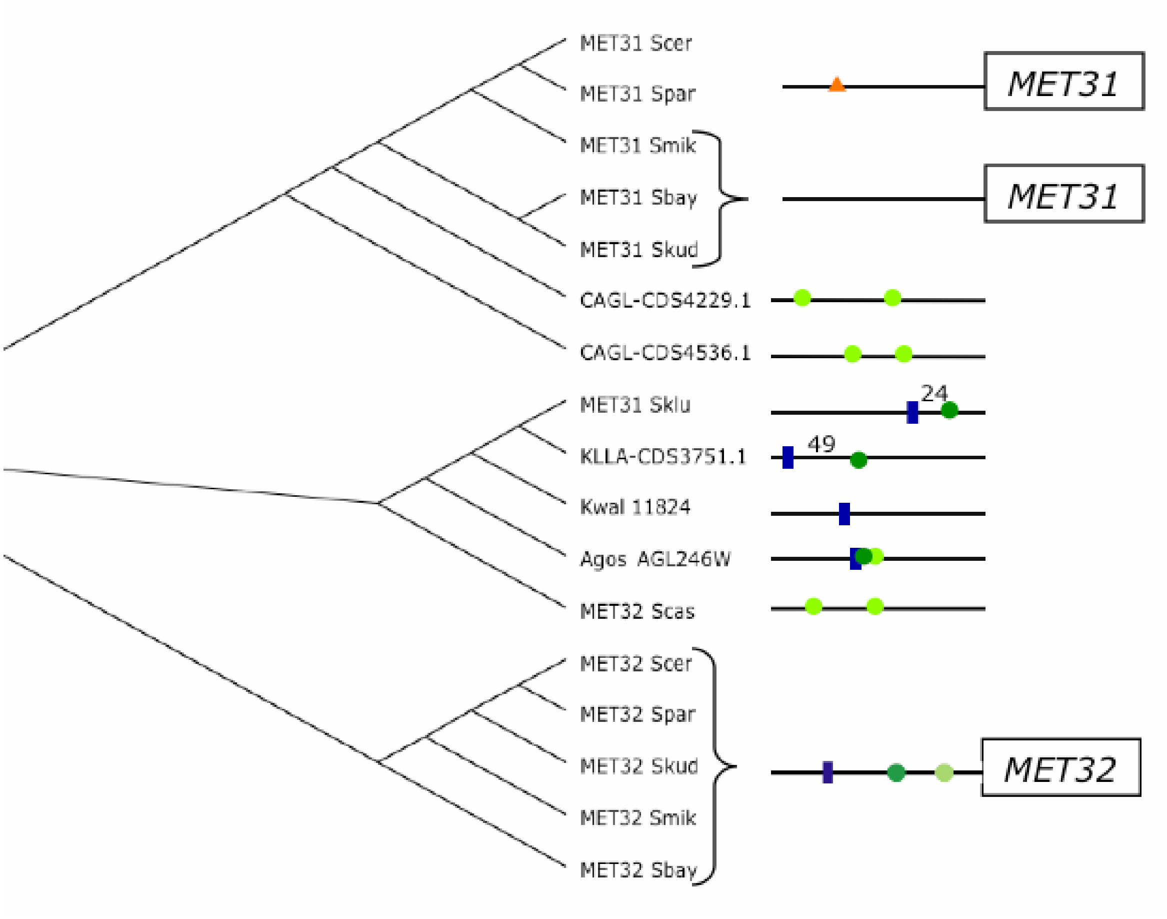
Differential regulatory roles ascribed to Met31 and Met32

Based on the above experimental data, I hypothesize that Met31 is primarily responsible for the cell cycle regulation of sulfur utilization genes, and that Met32 is activated only under sulfur-limiting conditions. Since Met28 stabilizes Cbf1 binding, Met32's ability to interact with Met28 suggests that it may recruit Met4 more potently than Met31 could. Notably, the promoters of single-copy homologs of *MET31* and *MET32* in *Saccharomyces kluyveri*, *Kluyveromyces lactis* and *Ashbya gossypii* also contain sequence motifs recognized by Cbf1 and Met31 or Met32 from *S. cerevisiae* (Figure 5.4). These shared binding sites suggest that the ancestral homolog of the Met31 or Met32 transcription factor was sulfur-regulated. The lack of accelerated evolution and the presence of Cbf1 and Met31 or Met32-like binding sites in the promoters single-copy homologs suggest that Met31 and Met32 are regulated under complementary conditions, in accordance with the subfunctionalization theory of Force *et al.*, 1999.

An experimental proposal to test theories on the evolution of duplicated genes

The accelerated evolution and duplication-divergence-complementation theories predict different outcomes for the ability of a single-copy ancestral homolog to complement a strain with mutations in both paralogs. Double mutant strains for transcription factor paralogs can be generated by crosses of the *Saccharomyces* deletion strain collection (Winzeler *et al.*, 1999). The outgroup homolog from either *K. waltii* or *A. gossypii*, along with its promoter, could be transformed into the corresponding double mutant strain and tested for its ability to complement the double mutant. If the

Figure 5.4) Putative regulatory sequences in promoters of Met31 or Met32 homologs



Homologs to *MET31* and *MET32* were assigned by genome sequencing projects. Putative binding sites in the 1000 bp upstream of translation start are indicated: Cbf1 (blue rectangles, TCACGTG), Met31 or Met32 (green circles, TGTGGC) and Gcn4 (orange triangles, TGA[C|G]TCA). The relative positions of binding sites are not drawn to scale. A phylogenetic tree was constructed from a T-COFFEE multiple alignment of amino acid sequences. Species abbreviations are as follows: *Saccharomyces cerevisiae* (Scer); *S. paradoxus* (Spar); *S. mikatae* (Smik); *S. kudriavzevii* (Skud); *S. bayanus* (Sbay); *S. castellii* (Scas); *S. kluyveri* (Sklu); *Candida glabrata* (CAGL); *Kluyveromyces lactis* (KLLA); *Kluyveromyces waltii* (Kwal); *Ashbya gossypii* (Agos).

subfunctionalization theory is correct, then the outgroup homolog should fully complement the double mutant. By contrast, if accelerated evolution has enabled the acquisition of a new function in one paralog, then the outgroup homolog would fail to fully complement the double mutant. It is possible that the amino acid sequences functionally complement, but that the transcriptional regulation of one paralog may have diverged. To test this possibility, these experiments could be repeated by integrating the outgroup homolog in the chromosomal locus of either *S. cerevisiae* paralog, thus placing it under the control of each derived promoter.

<u>CHAPTER 6</u>

CONCLUSIONS

This dissertation has focused on characterizing how promoter architecture governs multifactorial transcriptional regulation in yeast. Experimental studies on several mammalian enhanceosomes suggested that rigid distance constraints between transcription factor binding sites set the stereospecificity for the assembly of multiprotein regulatory complexes. Given the recent availability of multiple yeast genome sequences and genome-wide expression data, I investigated whether general principles on distance constraints between transcription factor binding sites could be inferred. In particular, I sought sequence-based rules that could predict whether any regulatory region could generate a particular condition-specific pattern of gene expression.

I have developed computational approaches that evaluated the transcriptional regulatory information encoded by DNA sequences of a fixed length. Genome-mean expression profiles indicated the regulatory potential of individual sequences. As a proof of principle, profiles generated from previously published microarray data could identify known transcription factor binding sites. Positional information for transcription factor binding sites was systematically confirmed, since the strongest associations with gene expression changes occurred for sequences found between 100 and 400 bp upstream of the translation start site. Notably, less than 20% of genes whose promoters contained a transcription factor binding site were associated with significant gene expression changes, thus suggesting the involvement of additional transcription factors.

By integrating comparative sequence data into the analysis of gene expression data, I confirmed the expectation that promoter architecture is under purifying selection. I predicted interactions between transcription factor binding sites using a series of statistical tests to identify pairs of DNA hexamers that were jointly conserved and closely spaced. Whereas computational analyses can often detect global trends, such as the importance of binding site spacing in transcription factor interactions, further experiments are often necessary to test predictions about individual examples. I am pleased to note that my predicted interaction between Hap1 and the pair, Ecm22p and Upc2, in the regulation of ergosterol biosynthesis has been experimentally confirmed by Jasper Rine's laboratory. Thus, my computational approaches to detect statistical associations between DNA sequences and changes in gene expression can successfully predict transcription factors involved in the multifactorial transcriptional regulation in yeast. These methods could be easily applied to find regulatory sequences in the core promoters of other eukaryotes whose genomes have been sequenced, and for which systematic gene expression data have been collected.

I conducted experiments in order to glean mechanistic insights on how promoter architecture influences the collaborative recruitment of the coactivator, Met4, by the yeast transcription factors, Cbf1p and Met31 or Met32p, under methionine starvation conditions. There were too few examples of annotated promoters available to make statistically significant inferences about helical phasing or distance constraints. Thus, I developed a synthetic promoter assay to assess the influence of varying distances and sequence contexts between these transcription factors. I confirmed that a Cbf1 binding site was required upstream of a Met31 or Met32 binding site for high levels of reporter gene activation *in vivo*. My results from growth rates of fixed-length promoter libraries revealed key differences in the distance constraints between transcription factor binding sites. In contrast with cooperatively binding transcription factors, whose synergistic activation decreases precipitously when the distance between binding sites was increased by more than 10 bp, levels of gene activation were fairly consistent when Cbf1 and Met31 or Met32 binding sites were spaced between 18 bp and 50 bp apart. Interestingly, binding sites spaced less than 18 bp apart could not support high levels of reporter gene activation, even though the individual binding sites did not overlap. I also discovered that certain sequence contexts dramatically diminished reporter gene activation.

Taken together, these experiments suggest that the process of Met4 recruitment differs considerably from the lock-and-key arrangements of bound transcription factors predicted by the enhanceosome model. Instead, DNA bendability may enforce an induced fit between the bound transcription factors and Met4. Whereas the distance between binding sites plays a diminished role in bridging bound transcription factors, intervening sequences with low intrinsic bendability could impair coactivator recruitment. Thus, the key requirements of promoter architecture may rely heavily on the molecular mechanism of transcription factor interactions at a particular set of co-regulated promoters. If promoter architectures were indeed idiosyncratic, it would be difficult to generalize experimental characterizations of transcription factor interactions. In addition, it would be very difficult to make highly accurate computational predictions of multifactorial regulation for individual regulatory regions, without some experimental knowledge about the interaction mechanisms of the relevant transcription factors.

The myriad effects of long-range interactions among transcription factor binding sites on multifactorial regulation pose similar challenges to predicting protein structures from secondary structure elements. Whereas general principles – such as close spacing of binding sites or van der Waals interactions among protein side chains – may govern these processes, they provide little predictive power for individual examples. Further studies on promoter architecture may benefit from a framework that formalizes how enthalpy gains from protein-protein interactions are offset by the entropy loss of multiprotein complex formation. Thermodynamic measurements on promoter variants with different spacing and sequence contexts between transcription factor binding sites could then be associated with changes in gene activation. Such a theory on the energetics of multiprotein complex formation could provide the quantitative precision needed to predict how a particular transcriptional control region adopts a conformation that enables transcriptional activation.

REFERENCES

Akache, B., K. Wu and B. Turcotte (2001). Phenotypic analysis of genes encoding yeast zinc cluster proteins. *Nucleic Acids Research* **29**: 2181-2190.

Amin, J., M. Fernandez, J. Ananthan, J. T. Lis and R. Voellmy (1994). Cooperative binding of heat shock transcription factor to the Hsp70 promoter in vivo and in vitro. *Journal of Biological Chemistry* **269**: 4804-4811.

Anderson, J. D., P. T. Lowary and J. Widom (2001). Effects of histone acetylation on the equilibrium accessibility of nucleosomal DNA target sites. *Journal of Molecular Biology* **307**: 977-985.

Aparicio, S., A. Morrison, A. Gould, J. Gilthorpe, C. Chaudhuri, P. Rigby, R. Krumlauf and S. Brenner (1995). Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, *Fugu rubripes*. *Proceedings of the National Academy of Sciences of the United States of America* **92**: 1684-1688.

Bailey, T. L. and C. Elkan (1995). Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning* **21**: 51-80.

Balakrishnan, R., K. R. Christie, M. C. Costanzo, K. Dolinski, S. S. Dwight, S. R. Engel, D. G. Fisk, J. E. Hirschman, E. L. Hong, R. Nash, R. Oughtred, M. Skrzypek, C. L. Theesfeld, G. Binkley, Q. Dong, C. Lane, A. Sethuraman, S. Weng, D. Botstein and J. M. Cherry (2005). Fungal BLAST and model organism BLASTP best hits: new comparison resources at the Saccharomyces Genome Database (SGD). *Nucleic Acids Research* **33**: D374-D377.

Barbaric, S., M. Munsterkotter, C. Coding and W. Horz (1998). Cooperative Pho2-Pho4 interactions at the *PHO5* promoter are critical for binding of Pho4 to UASp1 and for efficient transactivation by Pho4 and UASp2. *Molecular and Cellular Biology* **18**: 2629-2639.

Basehoar, A. D., S. J. Zanton and B. F. Pugh (2004). Identification and distinct regulation of yeast TATA box-containing genes. *Cell* **116**: 699-709.

Bateman, A., L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. L. Sonnhammer, D. J. Studholme, C. Yeats and S. R. Eddy (2004). The Pfam protein families database. *Nucleic Acids Research* **32**: D138-D141.

Becker, A. and G. Theissen (2003). The major clades of MADS-box genes and their role in the development and evolution of flowering plants. *Molecular Phylogenetics and Evolution* **29**: 464-489.

Becker, P. B. (2002). Nucleosome sliding: facts and fiction. *EMBO Journal* **21**: 4749-4753.

Beer, M. A. and S. Tavazoie (2004). Predicting gene expression from sequence. *Cell* **117**: 185-198.

Bender, A. and G. F. Sprague (1987). MAT alpha1 protein, a yeast transcription activator, binds synergistically with a second protein to a set of cell-type-specific genes. *Cell* **50**: 681-691.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* **57**: 289-300.

Berg, O. G. and P. H. von Hippel (1987). Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *Journal of Molecular Biology* **193**: 723-750.

Berman, B. P., Y. Nibu, B. D. Pfeiffer, P. Tomancek, S. E. Celniker, M. Levine, G. M. Rubin and M. B. Eisen (2002). Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proceedings of the National Academy of Sciences USA* **99**: 757-762.

Bernstein, B. E., C. L. Liu, E. L. Humphrey, E. O. Perlstein and S. L. Schreiber (2004). Global nucleosome occupancy in yeast. *Genome Biology* **5**.

Bhoite, L. T., Y. Yu and D. J. Stillman (2001). The Swi5 activator recruits the Mediator complex to the HO promoter without RNA polymerase II. *Genes & Development* **15**: 2457-2469.

Bhoite, L. T., J. M. Allen, E. Garcia, L. R. Thomas, I. D. Gregory, W. P. Voth, K. Whelihan, R. J. Rolfes and D. J. Stillman (2002). Mutations in the Pho2 (Bas2) transcription factor that differentially affect activation with its partner proteins Bas1, Pho4, and Swi5. *Journal of Biological Chemistry* **277**: 37612-37618.

Blackwood, E. M. and J. T. Kadonaga (1998). Going the distance: a current view of enhancer action. *Science* **281**: 60-63.

Blaiseau, P. L., A. D. Isnard, Y. Surdin-Kerjan and D. Thomas (1997). Met31p and Met32p, two related zinc finger proteins, are involved in transcriptional regulation of yeast sulfur amino acid metabolism. *Molecular and Cellular Biology* **17**: 3640-3648.

Blanchette, M. and M. Tompa (2002). Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Research* **12**: 739-748.

Boeger, H., J. Grisenbeck, J. S. Strattan and R. D. Kornberg (2003). Nucleosomes unfold completely at a transcriptionally active promoter. *Molecular Cell* **11**: 1587-1598.

Boffelli, D., J. McAuliffe, D. Ovcharenko, K. D. Lewis, I. Ovcharenko, L. Pachter and E. M. Rubin (2003). Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**: 1391-1394.

Bond, G. L., W. Hu, E. E. Bond, H. Robins, S. G. Lutzker, N. C. Arva, J. Bargonetti, F. Bartel, H. Taubert, P. Wuerl, K. Onel, L. Yip, S. L. Hwang, L. C. Strong, G. Lozano and A. J. Levine (2004). A single nucleotide polymorphism in the MDM2 promoter attenuates the p53 tumor suppressor pathway and accelerates tumor formation in humans. *Cell* **119**: 591-602.

Boros, J., F. L. Lim, Z. Darieva, A. Pic-Taylor, R. Harman, B. A. Morgan and A. D. Sharrocks (2003). Molecular determinants of the cell-cycle regulated Mcm1p-Fkh2p transcription factor complex. *Nucleic Acids Research* **31**: 2279-2288.

Brazas, R. M., L. T. Bhoite, M. D. Murphy, Y. Yu, Y. Chen, D. W. Neklason and D. J. Stillman (1995). Determining the requirements for cooperative DNA binding by Swi5p and Pho2p (Grf10p/Bas2p) at the *HO* promoter. *Journal of Biological Chemistry* **270**: 29151-29161.

Brazma, A., I. Jonassen, J. Vilo and E. Ukkonen (1998). Predicting gene regulatory elements in silico on a genomic scale. *Genome Research* **8**: 1202-1215.

Britten, R. J. and E. H. Davidson (1969). Gene regulation for higher cells: a theory. *Science* **165**: 349-357.

Brown, P. O. and D. Botstein (1999). Exploring the new world of the genome with DNA microarrays. *Nature Genetics* **21**: 33-37.

Bruhn, L. and G. F. Sprague (1994). MCM1 point mutants deficient in expression of alpha-specific genes: residues important for interaction with alpha 1. *Molecular and Cellular Biology* **14**: 2534-2544.

Bryant, G. O. and M. Ptashne (2003). Independent recruitment *in vivo* by Gal4 of two complexes required for transcription. *Molecular Cell* **11**: 1301-1309.

Bussemaker, H. J., H. Li and E. D. Siggia (2000). Building a dictionary for genomes: Identification of presumptive regulatory sites by statistical analysis. *Proceedings of the National Academy of Sciences of the United States of America* **97**: 10096-10100.

Bussemaker, H. J., H. Li and E. D. Siggia (2001). Regulatory element detection using correlation with expression. *Nature Genetics* **27**: 167-171.

Cai, M. and R. W. Davis (1990). The yeast centromere binding protein CBF1, of the helix-loop-helix protein family, is required for chromosome stability and methionine prototrophy. *Cell* **61**: 437-446.

Carey, M. (1998). The enhanceosome and transcriptional synergy. *Cell* **92**: 5-8.

Carr, E. A., J. Mead and A. K. Vershon (2004). α1-induced DNA bending is required for transcriptional activation by the Mcm1-α1 complex. *Nucleic Acids Research* **32**: 2298-2305.

Carrozza, M. J., R. T. Utley, J. L. Workman and J. Cote (2003). The diverse functions of histone acetyltransferase complexes. *Trends in Genetics* **19**: 321-329.

Causton, H. C., B. Ren, S. S. Koh, C. T. Harbison, E. Kanin, E. G. Jennings, T. I. Lee, H. L. True, E. S. Lander and R. A. Young (2001). Remodeling of yeast genome expression in response to environmental changes. *Molecular Biology of the Cell* **12**: 323-337.

Chatterjee, S. and K. Struhl (1995). Connecting a promoter-bound protein to TBP bypasses the need for a transcriptional activation domain. *Nature* **374**: 820-822.

Chen, T. and C. S. Parker (2002). Dynamic association of transcriptional activation domains and regulatory regions in *Saccharomyces cerevisiae* heat shock factor. *Proceedings of the National Academy of Sciences of the United States of America* **99**: 1200-1205.

Cherest, H. and Y. Surdin-Kerjan (1992). Genetic analysis of a new mutation conferring cysteine auxotrophy in *Saccharomyces cerevisiae*: updating of the sulfur metabolism pathway. *Genetics* **130**: 51-58.

Cherest, H., J. C. Davidian, D. Thomas, V. Benes, W. Ansorge and Y. Surdin-Kerjan (1997). Molecular characterization of two high affinity sulfate transporters in *Saccharomyces cerevisiae*. *Genetics* **145**: 627-635.

Chiang, D. Y., P. O. Brown and M. B. Eisen (2001). Visualizing associations between genome sequences and gene expression data using genome-mean expression profiles. *Bioinformatics* **17**: S49-S55.

Cho, R. J., M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart and R. W. Davis (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell* **2**: 65-73.

Chu, S., J. DeRisi, M. B. Eisen, J. Mulholland, D. Botstein, P. O. Brown and I. Herskowitz (1998). The transcriptional program of sporulation in budding yeast. *Science* **282**: 699-705.

Cliften, P. F., L. W. Hillier, L. Fulton, T. Graves, T. Miner, W. R. Gish, R. H. Waterston and M. Johnston (2001). Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis. *Genome Research* **11**: 1175-1186.

Cliften, P. F., P. Sudarsanam, A. Desikan, L. Fulton, B. Fulton, J. Majors, R. H. Waterston, B. A. Cohen and M. Johnston (2003). Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301**: 71-76.

Cohen, R. S. and M. Meselson (1988). Periodic interactions of heat-shock transcriptional elements. *Nature* **332**: 856-858.

Conlon, E. M., X. S. Liu, J. D. Lieb and J. S. Liu (2003). Integrating regulatory motif discovery and genome-wide expression analysis. *Proceedings of the National Academy of Sciences USA* **100**: 3339-3344.

Cooper, G. M. and A. Sidow (2003). Genomic regulatory regions: insights from comparative sequence analysis. *Current Opinion in Genetics & Development* **13**: 604-610.

Cosma, M. P., S. Panizza and K. Nasmyth (2001). Cdk1 triggers association of RNA polymerase to cell cycle promoters only after recruitment of the mediator by SBF. *Molecular Cell* **7**: 1213-1220.

Costanzo, M. C., M. E. Crawford, J. E. Hirschman, J. E. Kranz, P. Olsen, L. S. Robertson, M. S. Skrzypek, B. R. Braun, K. L. Hopkins, P. Kondu, C. Lengieza, J. E. Lew-Smith, M. Tillberg and J. I. Garrels (2001). YPD^{TM}, PombePD^{TM}, and WormPD^{TM}: model organism volumes of the BioKnowledge$^{®}$ library, an integrated resource for protein information. *Nucleic Acids Research* **29**: 75-79.

Crooks, G. E., G. Hon, J. M. Chandonia and S. E. Brenner (2004). WebLogo: A sequence logo generator. *Genome Research* **14**: 1188-1190.

Danilition, S. L., R. M. Frederickson, C. Y. Taylor and N. G. Miyamoto (1991). Transcription factor binding and spacing constraints in the human beta-actin proximal promoter. *Nucleic Acids Research* **19**: 6913-6922.

Davidson, E. H., J. P. Rast, P. Oliveri, A. Ransick, C. Calestani, C. H. Yuh, T. Minokawa, G. Amore, V. Hinman, C. Arenas-Mena, O. Otim, C. T. Brown, C. B. Livi, P. Y. Lee, R. Revilla, A. G. Rust, Z. J. Pan, M. J. Schilstra, P. J. C. Clarke,

M. I. Arnone, L. Rowen, R. A. Cameron, D. R. McClay, L. Hood and H. Bolouri (2002). A genomic regulatory network for development. *Science* **295**: 1669-1678.

Deckert, J. and K. Struhl (2001). Histone acetylation at promoters is differentially affected by specific activators and repressors. *Molecular and Cellular Biology* **21**: 2726-2735.

DeRisi, J., V. R. Iyer and P. O. Brown (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**: 680-686.

Diamond, M. I., J. N. Miner, S. K. Yoshinaga and K. R. Yamamoto (1990). Transcription factor interactions: selectors of positive or negative regulation from a single DNA element. *Science* **249**: 1266-1272.

Dietrich, F. S., S. Voegeli, S. Brachat, A. Lerch, K. Gates, S. Steiner, C. Mohr, R. Pohlmann, P. Luedi, S. D. Choi, R. A. Wing, A. Flavier, T. D. Gaffney and P. Phillippsen (2004). The Ashbya gossypii genome as a tool for mapping the ancient Saccharomyces cerevisiae genome. *Science* **304**: 304-307.

Dodou, E. and R. Treisman (1997). The Saccharomyces cerevisiae MADS-box transcription factor Rlm1 is a target for the Mpk1 mitogen-activated protein kinase pathway. *Molecular and Cellular Biology* **17**: 1848-1859.

Dohrmann, P. R., W. P. Voth and D. J. Stillman (1996). Role of negative regulation in promoter specificity of the homologous transcriptional activators Ace2p and Swi5p. *Molecular and Cellular Biology* **16**: 1746-1758.

Dormer, U. H., J. Westwater, N. F. McLaren, N. A. Kent, J. Mellor and D. J. Jamieson (2000). Cadmium-inducible expression of the yeast GSH1 gene requires a functional sulfur-amino acid regulatory network. *Journal of Biological Chemistry* **275**: 32611-32616.

Drazinic, C. M., J. B. Smerage, M. C. Lopez and H. V. Baker (1996). Activation mechanism of the multifunctional transcription factor repressor-activator protein 1 (Rap1p). *Molecular and Cellular Biology* **16**: 3187-3196.

Dueber, J. E., B. J. Yeh, R. P. Bhattacharyya and W. A. Lim (2004). Rewiring cell signaling: the logic and plasticity of eukaryotic protein circuitry. *Current Opinion in Structural Biology* **14**: 690-699.

Duffy, J. B. and N. Perrimon (1996). Recent advances in understanding signal transduction pathways in worms and flies. *Current Opinion in Cell Biology* **8**: 231-238.

Eisen, M. B., P. T. Spellman, P. O. Brown and D. Botstein (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences USA* **95**: 14863-14868.

Elledge, S. J. and R. W. Davis (1989). Position and density effects on repression by stationary and mobile DNA-binding proteins. *Genes & Development* **3**: 185-197.

Erives, A. and M. Levine (2004). Coordinate enhancers share common organizational features in the *Drosophila* genome. *Proceedings of the National Academy of Sciences of the United States of America* **101**: 3851-3856.

Escriva, H., R. Safi, C. Hanni, M. C. Langlois, P. SaumitouLaprade, D. Stehelin, A. Capron, R. Pierce and V. Laudet (1997). Ligand binding was acquired during evolution of nuclear receptors. *Proceedings of the National Academy of Sciences of the United States of America* **94**: 6803-6808.

Eskin, E. and P. A. Pevzner (2002). Finding composite regulatory patterns in DNA sequences. *Bioinformatics* **18**: S354-S363.

Falvo, J. V., D. Thanos and T. Maniatis (1995). Reversal of intrinsic DNA bends in the IFNb gene enhancer by transcription factors and the architectural protein HMG I(Y). *Cell* **83**.

Falvo, J. V., B. S. Parekh, C. H. Lin, E. Fraenkel and T. Maniatis (2000). Assembly of a functional beta interferon enhanceosome is dependent on ATF-2--c-jun heterodimer orientation. *Molecular and Cellular Biology* **20**: 4814-4825.

Flynn, P. J. and R. J. Reece (1999). Activation of transcription by metabolic intermediates of the pyrimidine biosynthetic pathway. *Molecular and Cellular Biology* **19**: 882-888.

Force, A., M. Lynch, F. B. Pickett, A. Amores, Y. Yan and J. Postlethwait (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**: 1531-1545.

Galas, D. J. and A. Schmitz (1978). DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Research* **5**: 3157-3170.

Garner, M. M. and A. Revzin (1981). A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the *Escherichia coli* lactose operon regulatory system. *Nucleic Acids Research* **9**: 3047-3060.

Gasch, A. P., P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein and P. O. Brown (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell* **11**: 4241-4257.

Gasch, A. P., M. Huang, S. Metzner, D. Botstein, S. J. Elledge and P. O. Brown (2001). Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p. *Molecular Biology of the Cell* **12**: 2987-3003.

Gasch, A. P. (2003). The environmental stress response: a common yeast response to diverse environmental stresses. Yeast Stress Responses. S. Hohmann and W. H. Mager. Berlin, Springer. **1:** 11-70.

Gasch, A. P., A. M. Moses, D. Y. Chiang, H. B. Fraser, M. Berardini and M. B. Eisen (2004). Conservation and evolution of cis-regulatory systems in ascomycete fungi. *Plos Biology* **2**: 2202-2219.

Gietz, R. D. and R. A. Woods (2002). Transformation of yeast by lithium acetate/single-stranded carrier DNA/polyethylene glycol method. *Methods in Enzymology* **350**: 87-96.

Giniger, E. and M. Ptashne (1988). Cooperative DNA-binding of the yeast transcriptional activator Gal4. *Proceedings of the National Academy of Sciences of the United States of America* **85**: 382-386.

Goto, N. K., T. Zor, M. Martinez-Yamout, H. J. Dyson and P. E. Wright (2002). Cooperativity in transcription factor binding to the coactivator CREB-binding protein (CBP). The mixed lineage leukemia protein (MLL) activation domain binds to an allosteric site on the KIX domain. *Journal of Biological Chemistry* **277**: 43168-43174.

Gregory, P. D., A. Schmid, M. Zavari, M. Munsterkotter and W. Horz (1999). Chromatin remodelling at the PHO8 promoter requires SWI-SNF and SAGA at a step subsequent to activator binding. *EMBO Journal* **18**: 6407-6414.

Grosschedl, R. (1995). Higher-order nucleoprotein complexes in transcription: analogies with site-specific recombination. *Current Opinion in Cell Biology* **7**: 362-370.

Gueldener, U., J. Heinisch, G. J. Koehler, D. Voss and J. H. Hegemann (2002). A second set of *loxP* marker cassettes for Cre-mediated multiple gene knockouts in budding yeast. *Nucleic Acids Research* **30**: e23.

Guet, C. C., M. B. Elowitz, W. H. Hsing and S. Leibler (2002). Combinatorial synthesis of genetic networks. *Science* **296**: 1466-1470.

GuhaThakurta, D. and G. D. Stormo (2001). Identifying target sites for cooperatively binding factors. *Bioinformatics* **17**: 608-621.

Gumucio, D. L., D. A. Shelton, W. Zhu, D. Millinoff, T. Gray, J. H. Bock, J. L. Slightom and M. Goodman (1996). Evolutionary strategies for the elucidation of *cis* and *trans* factors that regulate the developmental switching programs of the β-like globin genes. *Molecular Phylogenetics and Evolution* **5**: 18-32.

Hampsey, M. (1998). Molecular genetics of the RNA polymerase II general transcriptional machinery. *Microbiology and Molecular Biology Reviews* **62**: 465-503.

Hampsey, M. and D. Reinberg (1999). RNA polymerase II as a control panel for multiple coactivator complexes. *Current Opinion in Genetics & Development* **9**: 132-139.

Hanes, S. D., G. Riddihough, D. Ish-Horowicz and R. Brent (1994). Specific DNA recognition and intersite spacing are critical for action of the bicoid morphogen. *Molecular and Cellular Biology* **14**: 3364-3375.

Hannenhalli, S. and S. Levy (2002). Predicting transcription factor synergism. *Nucleic Acids Research* **30**: 4278-4284.

Herschlag, D. and F. B. Johnson (1993). Synergism in transcriptional activation: a kinetic view. *Genes & Development* **7**: 173-179.

Herskowitz, I. (1989). A regulatory hierarchy for cell specialization in yeast. *Nature* **342**: 749-757.

Hertz, G. Z. and G. D. Stormo (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15**: 563-577.

Hittinger, C. T., A. Rokas and S. B. Carroll (2004). Parallel inactivation of multiple *GAL* pathway genes and ecological diversification in yeasts. *Proceedings of the National Academy of Sciences of the United States of America* **101**: 14144-14149.

Ho, Y., A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S. L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Y. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. R. Willems, H. Sassi, P. A. Nielsen, K. J. Rasmussen, J. R. Andersen, L. E. Johansen, L. H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. D. Sorensen, J. Matthiesen, R. C. Hendrickson, F. Gleeson, T. Pawson, M. F. Moran, D. Durocher, M. Mann, C. W. V. Hogue, D. Figeys and M. Tyers (2002). Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**: 180-183.

Hollenhorst, P. C., M. E. Bose, M. R. Mielke, U. Muller and C. A. Fox (2000). Forkhead genes in transcriptional silencing, cell morphology and the cell cycle:

Overlapping and distinct functions for FKH1 and FKH2 in Saccharomyces cerevisiae. *Genetics* **154**: 1533-1548.

Holmes, I. and W. J. Bruno (2000). Finding regulatory elements using joint likelihoods for sequence and expression profile data. *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology* **8**: 202-210.

Horecka, J. and G. F. J. Sprague (2000). Use of imadazoleglyercerolphosphate dehydratase (His3) as a biological reporter in yeast. *Methods in Enzymology* **326**: 107-119.

Hsia, C. C. and W. McGinnis (2003). Evolution of transcription factor function. *Current Opinion in Genetics & Development* **13**: 199-206.

Hughes, J. D., P. W. Estep, S. Tavazoie and G. M. Church (2000a). Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *Journal of Molecular Biology* **296**: 1205-1214.

Hughes, T. R., M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Y. Dai, Y. D. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard and S. H. Friend (2000b). Functional discovery via a compendium of expression profiles. *Cell* **102**: 109-126.

Ikeda, K., Y. Watanabe, H. Ohto and K. Kawakami (2002). Molecular interaction and synergistic activation of a promoter by Six, Eya, and Dach proteins mediated through CREB binding protein. *Molecular and Cellular Biology* **22**: 6759-6766.

Inokuchi, K., A. Nakayama and F. Hishinuma (1987). Identification of sequence elements that confer cell-type-specific control of MF alpha 1 expression in *Saccharomyces cerevisiae*. *Molecular and Cellular Biology* **7**: 3185-3193.

Ioshikhes, I., E. N. Trifonov and M. Q. Zhang (1999). Periodical distribution of transcription factor sites in promoter regions and connection with chromatin structure. *Proceedings of the National Academy of Sciences of the United States of America* **96**: 2891-2895.

Iyer, V. R., C. E. Horak, C. S. Scafe, D. Botstein, M. Snyder and P. O. Brown (2001). Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**: 533-538.

Jarvis, E. E., D. C. Hagen and G. F. Sprague (1988). Identification of a DNA segment that is necessary and sufficient for alpha-specific gene control in *Saccharomyces cerevisiae*: implications for regulation of alpha-specific and a-specific genes. *Molecular and Cellular Biology* **8**: 309-320.

Jin, Y., J. Mead, T. Li, C. Wolberger and A. K. Vershon (1995). Altered DNA recognition and bending by insertions in the α2 tail of the yeast a1/α2 homeodomain heterodimer. *Science* **270**: 290-292.

Johnson, A. D. (1995). Molecular mechanisms of cell-type determination in budding yeast. *Current Opinion in Genetics & Development* **5**: 552-558.

Kaffman, A., N. M. Rank, E. M. O'Neill, L. S. Huang and E. K. O'Shea (1998). The receptor Msn5 exports the phosphorylated transcription factor Pho4 out of the nucleus. *Nature* **396**.

Kaiser, P., K. Flick, C. Wittenberg and S. I. Reed (2000). Regulation of transcription by ubiquitination without proteolysis: Cdc34/SCF(Met30)-mediated inactivation of the transcription factor Met4. *Cell* **102**: 303-314.

Keegan, L., G. Gill and M. Ptashne (1986). Separation of DNA-binding from the transcription-activating function of a eukaryotic regulatory protein. *Science* **231**: 699-704.

Kel-Margoulis, O. V., T. G. Ivanova, E. Wingender and A. E. Kel (2002a). Automatic annotation of genomic regulatory sequences by searching for composite clusters. *Pac Symp Biocomput* **7**: 187-198.

Kel-Margoulis, O. V., A. E. Kel, I. Reuter, I. V. Deineko and E. Wingender (2002b). TRANSCompel((R)): a database on composite regulatory elements in eukaryotic genes. *Nucleic Acids Research* **30**: 332-334.

Keles, S., M. van der Laan and M. B. Eisen (2002). Identification of regulatory elements using a feature selection method. *Bioinformatics* **18**: 1167-1175.

Kellis, M., N. Patterson, M. Endrizzi, B. Birren and E. S. Lander (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**: 241-254.

Kellis, M., B. W. Birren and E. S. Lander (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature* **428**: 617-624.

Kent, N. A., S. M. Eibert and J. Mellor (2004). Cbf1p is required for chromatin remodeling at promoter-proximal CACGTG motifs in yeast. *Journal of Biological Chemistry* **279**: 27116-27123.

Kim, T. K., S. Hashimoto, R. J. Kelleher, P. M. Flanagan, R. D. Kornberg, M. Horikoshi and R. G. Roeder (1994). Effects of activation-defective TBP mutations on transcription initiation in yeast. *Nature* **369**: 252-255.

Kim, T. K. and T. Maniatis (1997). The mechanism of transcriptional synergy of an in vitro assembled interferon-β enhanceosome. *Molecular Cell* **1**: 119-129.

Kimura, A., T. Umehara and M. Horikoshi (2002). Chromosomal gradient of histone acetylation established by Sas2p and Sir2p functions as a shield against gene silencing. *Nature Genetics* **32**: 370-377.

King, D. A., L. Zhang, L. Guarente and R. Marmorstein (1999). Structure of a HAP1-DNA complex reveals dramatically asymmetric DNA binding by a homodimeric protein. *Nature Structural Biology* **6**: 64-71.

Klages, N. and M. Strubin (1995). Stimulation of RNA polymerase II transcription initiation by recruitment of TBP *in vivo*. *Nature* **374**: 822-823.

Klingenhoff, A., K. Frech, K. Quandt and T. Werner (1999). Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity. *Bioinformatics* **15**: 180-186.

Kuras, L., R. Barbey and D. Thomas (1997). Assembly of a bZIP-bHLH transcription activation complex: Formation of the yeast Cbf1-Met4-Met28 complex is regulated through Met28 stimulation of Cbf1 DNA binding. *EMBO Journal* **16**: 2441-2451.

Kuras, L. and K. Struhl (1999). Binding of TBP to promoters *in vivo* is stimulated by activators and requires Pol II holoenzyme. *Nature* **399**: 609-613.

Kuras, L., P. Kosa, M. Mencia and K. Struhl (2000). TAF-containing and TAF-independent forms of transcriptionally active TBP *in vivo*. *Science* **288**: 1244-1248.

Kuras, L., T. Borggrefe and R. D. Kornberg (2003). Association of the mediator complex with enhancers of active genes. *Proceedings of the National Academy of Sciences of the United States of America* **100**: 13887-13891.

Kwast, K. E., P. V. Burke and R. O. Poyton (1998). Oxygen sensing and the transcriptional regulation of oxygen-responsive genes in yeast. *Journal of Experimental Biology* **201**: 1177-1195.

Larschan, E. and F. Winston (2001). The *S. cerevisiae* SAGA complex functions *in vivo* as a coactivator for transcriptional activation by Gal4. *Genes & Development* **15**: 1946-1956.

Lawrence, C. E., S. F. Altschul, M. S. Boguski, J. S. Liu, F. Neuwald and J. C. Wootton (1993). Detecting subtle sequence signals - a Gibbs sampling strategy for multiple alignment. *Science* **262**: 208-214.

Lee, S. E., A. Pellecioli, J. Demeter, M. P. Vaze, A. P. Gasch, A. Malkova, P. O. Brown, D. Botstein, T. Stearns, M. Foiani and J. E. Haber (2000). Biological Responses to DNA Damage. Cold Spring Harbor, NY, Cold Spring Harbor Laboratory Press. **65:** 303-314.

Li, T., M. R. Stark, A. D. Johnson and C. Wolberger (1995). Crystal structure of the MATa1/MATα2 homeodomain heterodimer bound to DNA. *Science* **270**: 262-269.

Li, X. Y., A. Virbasius, X. C. Zhu and M. R. Green (1999). Enhancement of TBP binding by activators and general transcription factors. *Nature* **399**: 605-609.

Li, X. Y., S. R. Bhaumik and M. R. Green (2000). Distinct classes of yeast promoters revealed by differential TAF recruitment. *Science* **288**: 1242-1244.

Liao, G., J. Wang, J. Guo, J. Allard, J. Cheng, A. Ng, S. Shafer, A. Puech, J. D. McPherson, D. Foernzler, G. Peltz and J. Usuka (2004). In silico genetics: identification of a functional element regulating H2-Ealpha gene expression. *Science* **306**: 690-695.

Liao, G. C., E. J. Rehm and G. M. Rubin (2000). Insertion site preferences of the P transposable element in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences USA* **97**: 3347-3351.

Lim, F. L., A. Hayes, A. G. West, A. Pic-Taylor, Z. Darieva, B. A. Morgan, S. G. Oliver and A. D. Sharrocks (2003). Mcm1p-induced DNA bending regulates the formation of ternary transcription factor complexes. *Molecular and Cellular Biology* **23**: 450-461.

Liu, X., D. L. Brutlag and J. S. Liu (2001). BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pacific Symposium on Biocomputing* **6**: 127-38.

Lohr, U., M. Yussa and L. Pick (2001). *Drosophila fushi tarazu*: a gene on the border of homeotic function. *Current Biology* **11**: 1403-1412.

Loots, G. G., R. M. Locksley, C. M. Blankespoor, Z. E. Wang, W. Miller, E. M. Rubin and K. A. Frazer (2000). Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288**: 136-140.

Loots, G. G., I. Ovcharenko, L. Pachter, I. Dubchak and E. M. Rubin (2002). rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Research* **12**: 832-839.

Ludwig, M. Z., C. Bergman, N. H. Patel and M. Kreitman (2000). Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* **403**: 564-567.

Luger, K., A. W. Mader, R. K. Richmond, D. F. Sargent and T. J. Richmond (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**: 251-260.

Luscombe, N. M., S. E. Austin, H. M. Berman and J. M. Thornton (2000). An overview of the structures of protein-DNA complexes. *Genome Biology* **1**: reviews001.1-001.37.

Lusser, A. and J. T. Kadonaga (2003). Chromatin remodeling by ATP-dependent molecular machines. *BioEssays* **25**: 1192-1200.

Ma, J. and M. Ptashne (1987a). A new class of yeast transcriptional activators. *Cell* **51**: 113-119.

Ma, J. and M. Ptashne (1987b). The carboxy-terminal 20 amino acids of GAL4 are recognized by GAL80. *Cell* **50**: 137-142.

Mai, X., S. Chou and K. Struhl (2000). Preferential accessibility of the yeast *his3* promoter is determined by a general property of the DNA sequence, not by specific elements. *Molecular and Cellular Biology* **20**: 6668-6676.

Makeev, V. J., A. P. Lifanov, A. G. Nazina and D. A. Papatsenko (2003). Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information. *Nucleic Acids Research* **31**: 6016-6026.

Marsan, L. and M.-F. Sagot (2000). Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. *Journal of Computational Biology* **7**: 345-362.

Martens, J. A. and F. Winston (2003). Recent advances in understanding chromatin remodeling by Swi/Snf complexes. *Current Opinion in Genetics & Development* **13**: 136-142.

Matys, V., E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D. U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Munch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele and E. Wingender (2003). TRANSFAC (R): transcriptional regulation, from patterns to profiles. *Nucleic Acids Research* **31**: 374-378.

McAdams, H. H., B. Srinivasan and A. P. Arkin (2004). The evolution of genetic regulatory systems in bacteria. *Nature Reviews Genetics* **5**: 169-178.

Mead, J., H. Zhong, T. B. Acton and A. K. Vershon (1996). The yeast α2 and Mcm1 proteins interact through a region similar to a motif found in homeodomain proteins of higher eukaryotes. *Molecular and Cellular Biology* **16**: 2135-2143.

Mead, J., A. R. Bruning, M. K. Gill, A. M. Steiner, T. B. Acton and A. K. Vershon (2002). Interactions of the Mcm1 MADS box protein with cofactors that regulate mating in yeast. *Molecular and Cellular Biology* **22**: 4607-4621.

Melcher, K., B. Sharma, W. V. Ding and M. Nolden (2000). Zero background yeast reporter plasmids. *Gene* **247**: 53-61.

Merika, M., A. J. Williams, G. Chen, T. Collins and D. Thanos (1998). Recruitment of CBP/p300 by the IFNβ enhanceosome is required for synergistic activation of transcription. *Molecular Cell* **1**: 277-287.

Merika, M. and D. Thanos (2001). Enhanceosomes. *Current Opinion in Genetics and Development* **11**: 205-208.

Miller, J. A. and J. Widom (2003). Collaborative competition mechanism for gene activation in vivo. *Molecular and Cellular Biology* **23**: 1623-1632.

Mirny, L. A. and M. S. Gelfand (2002). Structural analysis of conserved base pairs in protein–DNA complexes. *Nucleic Acids Research* **30**: 1704-1711.

Moreau, J.-L., M. Lee, N. Mahachi, J. Vary, J. Mellor, T. Tsukiyama and C. R. Goding (2003). Regulated displacement of TBP from the *PHO8* promoter *in vivo* requires Cbf1 and the Isw1 chromatin remodeling complex. *Molecular Cell* **11**: 1609-1620.

Morillon, A., J. O'Sullivan, A. Azad, N. Proudfoot and J. Mellor (2003). Regulation of elongating RNA polymerase II by forkhead transcription factors in yeast. *Science* **300**: 492-495.

Moses, A. M., D. Y. Chiang, M. Kellis, E. S. Lander and M. B. Eisen (2003). Position specific variation in the rate of evolution in transcription factor binding sites. *BMC Evolutionary Biology* **3**: 19.1-19.13.

Moses, A. M., D. Y. Chiang and M. B. Eisen (2004). Phylogenetic motif detection by expectation-maximization on evolutionary mixtures. *Pacific Symposium on Biocomputing* **9**: 324-335.

Mueller, C. G. and A. Nordheim (1991). A protein domain conserved between yeast MCM1 and human SRF directs ternary complex formation. *EMBO Journal* **10**: 4219-4229.

Myers, L. C. and R. D. Kornberg (2000). Mediator of transcriptional regulation. *Annual Review of Biochemistry* **69**: 729-749.

Naar, A. M., B. D. Lemon and R. Tjian (2001). Transcriptional coactivator complexes. *Annual Review of Biochemistry* **70**: 475-501.

Natarajan, K., M. R. Meyer, B. M. Jackson, D. Slade, C. Roberts, A. G. Hinnebusch and M. J. Marton (2001). Transcriptional profiling shows that Gcn4p is a master regulator of gene expression during amino acid starvation in yeast. *Molecular and Cellular Biology* **21**: 4347-4368.

Niedenthal, R. K., M. Sen-Gupta, A. Wilmen and J. H. Hegemann (1993). Cpf1 protein induced bending of yeast centromere DNA element I. *Nucleic Acids Research* **21**: 4726-4733.

Notredame, C., D. G. Higgins and J. Heringa (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology* **302**: 205-217.

O'Conallain, C., M. T. Doolin, C. Taggart, F. Thornton and G. Butler (1999). Regulated nuclear localisation of the yeast transcription factor Ace2p controls expression of chitinase (CTS1) in Saccharomyces cerevisiae. *Molecular and General Genetics* **262**: 275-282.

Ogata, K., K. Sato and T. H. Tahirov (2003). Eukaryotic transcriptional regulatory complexes: cooperativity from near and afar. *Current Opinion in Structural Biology* **13**: 40-48.

Ogawa, N., J. DeRisi and P. O. Brown (2000). New components of a system for phosphate accumulation and polyphosphate metabolism in Saccharomyces cerevisiae revealed by genomic expression analysis. *Molecular Biology of the Cell* **11**: 4309-4321.

Ohler, U. and H. Niemann (2001). Identification and analysis of eukaryotic promoters: recent computational approaches. *Trends in Genetics* **17**: 56-60.

Ohno, S. (1970). Evolution by Gene Duplication. Heidelberg, Germany, Springer-Verlag.

Oliphant, A. R., C. J. Brandl and K. Struhl (1989). Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. *Molecular and Cellular Biology* **9**: 2944-2949.

Olson, W. K., A. A. Gorin, X.-J. Lu, L. M. Hock and V. B. Zhurkin (1998). DNA sequence-dependent deformability deduced from protein-DNA crystal complexes.

Proceedings of the National Academy of Sciences of the United States of America **95**: 11163-11168.

Patton, E. E., C. Peyraud, A. Rouillon, Y. Surdin-Kerjan, M. Tyers and D. Thomas (2000). SCFMet30-mediated control of the transcriptional activator Met4 is required for the G(1)-S transition. *EMBO Journal* **19**: 1613-1624.

Pavlidis, P., T. S. Furey, M. Liberto, D. Haussler and W. N. Grundy (2001). Promoter region-based classification of genes. *Proceedings of the Pacific Symposium on Biocomputing* **6**: 151-164.

Pearce, D., W. Matsui, J. N. Miner and K. R. Yamamoto (1998). Glucocorticoid receptor transcriptional activity determined by spacing of receptor and nonreceptor DNA sites. *Journal of Biological Chemistry* **273**: 30081-30085.

Pennacchio, L. A. and E. M. Rubin (2001). Genomic strategies to identify mammalian regulatory sequences. *Nature Reviews Genetics* **2**: 100-109.

Pilpel, Y., P. Sudarsanam and G. M. Church (2001). Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genetics* **29**: 153-159.

Polach, K. J. and J. Widom (1996). A model for the cooperative binding of eukaryotic regulatory proteins to nucleosomal target sites. *Journal of Molecular Biology* **258**: 800-812.

Pollard, D. A., C. Bergman, J. Stoye, S. E. Celniker and M. B. Eisen (2004). Benchmarking tools for the alignment of functional noncoding DNA. *BMC Bioinformatics* **5**: 6.

Pollock, R. and R. Treisman (1990). A sensitive method for the determination of protein-DNA binding specificities. *Nucleic Acids Research* **18**: 6197-6204.

Prakash, A., M. Blanchette, S. Sinha and M. Tompa (2004). Motif discovery in heterogeneous sequence data. *Pacific Symposium on Biocomputing* **9**: 348-359.

Press, W. H., S. A. Teukolsky, W. T. Vertterling and B. P. Flannery (1992). Numerical Recipes in C, Second Edition. Cambridge, Cambridge University Press.

Ptashne, M. (1988). How eukaryotic transcriptional activators work. *Nature* **335**: 683-689.

Ptashne, M. and A. Gann (1997). Transcriptional activation by recruitment. *Nature* **386**: 569-577.

Qiu, H. F., C. H. Hu, S. P. Yoon, K. Natarajan, M. J. Swanson and A. G. Hinnebusch (2004). An array of coactivators is required for optimal recruitment of TATA binding protein and RNA polymerase II by promoter-bound Gcn4p. *Molecular and Cellular Biology* **24**: 4104-4117.

Qiu, P., W. Ding, Y. Jiang, J. R. Greene and L. Wang (2002). Computational analysis of composite regulatory elements. *Mammalian Genome* **13**: 327-332.

Rastinejad, F., T. Perlmann, R. M. Evans and P. B. Sigler (1995). Structural determinants of nuclear receptor assembly on DNA direct repeats. *Nature* **375**: 203-211.

Reece, R. J. and M. Ptashne (1993). Determinants of binding-site specificity among yeast C_6 zinc cluster proteins. *Science* **261**: 909-911.

Reid, J. L., V. R. Iyer, P. O. Brown and K. Struhl (2000). Coordinate regulation of yeast ribosomal protein genes is associated with targeted recruitment of Esa1 histone acetylase. *Molecular Cell* **6**: 1297-1307.

Reinke, H. and W. Horz (2003). Histones are first hyperacetylated and then lose contact with the activated *PHO5* promoter. *Molecular Cell* **11**: 1599-1607.

Remenyi, A., H. R. Scholer and M. Wilmanns (2004). Combinatorial control of gene expression. *Nature Structural & Molecular Biology* **11**: 812-815.

Roberts, C. J., B. Nelson, M. J. Marton, R. Stoughton, M. R. Meyer, H. A. Bennett, Y. D. D. He, H. Y. Dai, W. L. Walker, T. R. Hughes, M. Tyers, C. Boone and S. H. Friend (2000). Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* **287**: 873-880.

Robinson, K. A. and J. M. Lopes (2000). *Saccharomyces cerevisiae* basic helix-loop-helix proteins regulate diverse biological processes. *Nucleic Acids Research* **28**: 1499-1505.

Rockman, M. V., M. W. Hahn, N. Soranzo, D. A. Loisel, D. B. Goldstein and G. A. Wray (2004). Positive selection on MMP3 regulation has shaped heart disease risk. *Current Biology* **14**: 1531-1539.

Rusch, J. and M. Levine (1996). Threshold responses to the dorsal regulatory gradient and the subdivision of primary tissue territories in the *Drosophila* embryo. *Current Opinion in Genetics & Development* **6**: 416-423.

Rustici, G., J. Mata, K. Kivinen, P. Lio, C. J. Penkett, G. Burns, J. Hayles, A. Brazma, P. Nurse and J. Bahler (2004). Periodic gene expression program of the fission yeast cell cycle. *Nature Genetics* **36**: 809-817.

Sandelin, A., W. Alkema, P. Engstrom, W. W. Wasserman and B. Lenhard (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research* **32**: D91-D94.

Schneider, T. D. and R. M. Stephens (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research* **18**: 6097-6100.

Schwartz, S., Z. Zhang, K. A. Frazer, A. Smit, C. Riemer, J. Bouck, R. Gibbs, R. Hardison and W. Miller (2000). PipMaker--a web server for aligning two genomic DNA sequences. *Genome Research* **10**: 577-586.

Sellick, C. A. and R. J. Reece (2003). Modulation of transcription factor function by an amino acid: activation of Put3p by proline. *EMBO Journal* **22**: 5147-5153.

Setty, Y., A. E. Mayo, M. G. Surette and U. Alon (2003). Detailed map of a cis-regulatory input function. *Proceedings of the National Academy of Sciences of the United States of America* **100**: 7702-7707.

Sherlock, G., T. Hernandez-Boussard, A. Kasarskis, G. Binkley, J. C. Matese, S. S. Dwight, M. Kaloper, S. Weng, H. Jin, C. A. Ball, M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein and J. M. Cherry (2001). The Stanford Microarray Database. *Nucleic Acids Research* **29**: 152-155.

Sikorski, R. S. and P. Hieter (1989). A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in *Saccharomyces cerevisiae*. *Genetics* **122**: 19-27.

Sinha, S. and M. Tompa (2000). A statistical method for finding transcription factor binding sites. *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology* **8**: 344-354.

Smith, D. L. and A. D. Johnson (1992). A molecular mechanism for combinatorial control in yeast: MCM1 protein sets the spacing and orientation of the homeodomains of an alpha 2 dimer. *Cell* **68**: 133-142.

Spellman, P. T., G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein and B. Futcher (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell* **9**: 3273-3297.

Spiegelman, B. M. and R. Heinrich (2004). Biological control through regulated transcriptional coactivators. *Cell* **119**: 157-167.

Stahl, B. D. and C. D. Allis (2000). The language of covalent histone modifications. *Nature* **403**: 41-45.

Stormo, G. D. (2000). DNA binding sites: representation and discovery. *Bioinformatics* **16**: 16-23.

Struhl, K. (1995). Yeast transcriptional regulatory mechanisms. *Annual Review of Genetics* **29**: 651-674.

Struhl, K. (1998). Histone acetylation and transcriptional regulatory mechanisms. *Genes & Development* **12**: 599-606.

Struhl, K. (1999). Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell* **98**: 1-4.

Sudarsanam, P., V. R. Iyer, P. O. Brown and F. Winston (2000). Whole-genome expression analysis of *snf/swi* mutants of *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences of the United States of America* **97**: 3364-3369.

Suka, N., K. Luo and M. Grunstein (2002). Sir2p and Sas2p opposingly regulate acetylation of yeast histone H4 lysine16 and spreading of heterochromatin. *Nature Genetics* **32**: 378-383.

Sun, K. M., E. Coic, Z. Q. Zhou, P. Durrens and J. E. Haber (2002). Saccharomyces forkhead protein Fkh1 regulates donor preference during mating-type switching through the recombination enhancer. *Genes & Development* **16**: 2085-2096.

Svaren, J. and W. Horz (1997). Transcription factors vs nucleosomes: regulation of the *PHO5* promoter in yeast. *Trends in Biochemical Sciences* **22**: 93-97.

Swanson, M. J., H. Qiu, L. Sumibcay, A. Krueger, S.-J. Kim, K. Natarajan, S. Yoon and A. G. Hinnebusch (2003). A multiplicity of coactivators is required by Gcn4p at individual promoter *in vivo*. *Molecular and Cellular Biology* **23**: 2800-2820.

Sze, J.-Y., M. Woontner, J. A. Jaehning and G. B. Kohlhaw (1992). *In vitro* transcriptional activation by a metabolic derivative: activation of Leu3 depends on alpha-isopropylmalate. *Science* **258**: 1142-1145.

Tagle, D. A., B. F. Koop, M. Goodman, J. L. Slightom, D. L. Hess and R. T. Jones (1988). Embryonic ε and γ globin genes of a prosimian primate (*Galago crassicaudatus*) nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *Journal of Molecular Biology* **203**: 439-455.

Tajima, F. (1993). Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics* **135**: 599-607.

Takahashi, K., M. Vigneron, H. Matthes, A. Wildeman, M. Zenke and P. Chambon (1986). Requirement of Stereospecific Alignments for Initiation from the Simian Virus-40 Early Promoter. *Nature* **319**: 121-126.

Tan, S. and T. J. Richmond (1998). Crystal structure of the yeast MATα2/MCM1/DNA ternary complex. *Nature* **391**: 660-666.

Tanaka, M. (1996). Modulation of promoter occupancy by cooperative DNA binding and activation-domain function is a major determinant of transcriptional regulation by activators *in vivo*. *Proceedings of the National Academy of Sciences of the United States of America* **93**: 4311-4315.

Tavazoie, S., J. D. Hughes, M. J. Campbell, R. J. Cho and G. M. Church (1999). Systematic determination of genetic network architecture. *Nature Genetics* **22**: 281-285.

Taverner, N. V., J. C. Smith and F. C. Wardle (2004). Identifying transcriptional targets. *Genome Biology* **5**: 210.

Thomas, D., H. Cherest and Y. Surdin-Kerjan (1989). Elements involved in *S*-adenosylmethionine mediated regulation of the *Saccharomyces cerevisiae MET25* gene. *Molecular and Cellular Biology* **9**: 3292-3298.

Thomas, D., I. Jacquemin and Y. Surdin-Kerjan (1992). Met4, a leucine zipper protein, and Centromere-binding factor-I are both required for transcriptional activation of sulfur metabolism in *Saccharomyces cerevisiae*. *Molecular and Cellular Biology* **12**: 1719-1727.

Thomas, D., L. Kuras, R. Barbey, H. Cherest, P. L. Blaiseau and Y. Surdin-Kerjan (1995). Met30, a yeast transcriptional inhibitor that responds to *S*-adenosylmethionine, is an essential protein with WD40 repeats. *Molecular and Cellular Biology* **15**: 6526-6534.

Thomas, D. and Y. Surdin-Kerjan (1997). Metabolism of sulfur amino acids in *Saccharomyces cerevisiae*. *Microbiology and Molecular Biology Reviews* **61**: 503-532.

Tsong, A. E., M. G. Miller, R. M. Raisner and A. D. Johnson (2003). Evolution of a combinatorial transcriptional circuit: A case study in yeasts. *Cell* **115**: 389-399.

van Helden, J., B. Andre and J. Collado-Vides (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *Journal of Molecular Biology* **281**: 827-842.

Vashee, S., K. Melcher, W. V. Ding, S. A. Johnston and T. Kodadek (1998). Evidence for two modes of cooperative DNA binding in vivo that do not involve direct protein-protein interactions. *Current Biology* **8**: 452-458.

Vershon, A. K. and M. Pierce (2000). Transcriptional regulation of meiosis in yeast. *Current Opinion in Cell Biology* **12**: 334-339.

Vik, A. and J. Rine (2001). Upc2p and Ecm22p, dual regulators of sterol biosynthesis in *Saccharomyces cerevisiae*. *Molecular and Cellular Biology* **21**: 6395-6405.

Vilo, J., A. Brazma, I. Jonassen, A. Robinson and E. Ukkonen (2000). Mining for putative regulatory elements in the yeast genome using gene expression data. *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology* **8**: 384-94.

Vogel, C., M. Bashton, N. D. Kerrison, C. Chothia and S. A. Teichmann (2004). Structure, function and evolution of multidomain proteins. *Current Opinion in Structural Biology* **14**: 208-216.

Wagner, A. (1999). Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics* **15**: 776-784.

Wang, T. and G. D. Stormo (2003). Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics* **19**: 2369-2380.

Wang, W., J. M. Cherry, D. Botstein and H. Li (2002). A systematic approach to reconstructing transcription networks in *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences USA* **99**: 16893-16898.

Wasserman, W. W. and J. W. Fickett (1998). Identification of regulatory regions which confer muscle-specific gene expression. *Journal of Molecular Biology* **278**: 167-181.

Werner, T., S. Fessele, H. Maier and P. J. Nelson (2003). Computer modeling of promoter organization as a tool to study transcriptional coregulation. *FASEB Journal* **17**: 1228-1237.

Williams, R. M., M. Primig, B. K. Washburn, E. A. Winzeler, M. Bellis, C. Sarrauste de Menthiere, R. W. Davis and R. E. Esposito (2002). The Ume6 regulon coordinates metabolic and meiotic gene expression in yeast. *Proceedings of the National Academy of Sciences USA* **99**: 13431-13436.

Winzeler, E. A., D. D. Shoemaker, A. Astromoff, H. Liang, K. Anderson, B. Andre, R. Bangham, R. Benito, J. D. Boeke, H. Bussey, A. M. Chu, C. Connelly, K. Davis, F. Dietrich, S. W. Dow, M. El Bakkoury, F. Foury, S. H. Friend, E. Gentalen, G. Giaever, J. H. Hegemann, T. Jones, M. Laub, H. Liao, N. Liebundguth, D. J. Lockhart, A. Lucau-Danila, M. Lussier, N. M'Rabet, P. Menard, M. Mittmann, C. Pai, C. Rebischung, J. L. Revuelta, L. Riles, C. J. Roberts, P. Ross-MacDonald, B. Scherens, M. Snyder, S. Sookhai-Mahadeo, R. K. Storms, S. Veronneau, M. Voet, G. Volckaert, T. R. Ward, R. Wysocki, G. S. Yen, K. X. Yu, K. Zimmermann, P. Philippsen, M. Johnston and R. W. Davis (1999). Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**: 901-906.

Wolberger, C., A. K. Vershon, B. Liu, A. D. Johnson and C. O. Pabo (1991). Crystal structure of a MAT α2 homeodomain-operator complex suggests a general model for homeodomain-DNA interactions. *Cell* **67**: 517-528.

Wolberger, C. (1999). Multiprotein-DNA complexes in transcriptional regulation. *Annual Review of Biophysics and Biomolecular Structure* **28**: 29-56.

Wolfsberg, T. G., A. E. Gabrielian, M. J. Campbell, R. J. Cho, J. L. Spouge and D. Landsman (1999). Candidate regulatory sequence elements for cell cycle-dependent transcription in Saccharomyces cerevisiae. *Genome Research* **9**: 775-792.

Wray, G. A., M. W. Hahn, E. Abouheif, J. P. Balhoff, M. Pizer, M. V. Rockman and L. A. Romano (2003). The evolution of transcriptional regulation in eukaryotes. *Molecular Biology and Evolution* **20**: 1377-1419.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences* **13**: 555-556.

Zhu, G. F., P. T. Spellman, T. Volpe, P. O. Brown, D. Botstein, T. N. Davis and B. Futcher (2000). Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth. *Nature* **406**: 90-94.

Zhu, J. and M. Q. Zhang (1999). SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics* **15**: 607-611.